

PAPER

A Generative Artificial Intelligence-Based Mechanism for Mobile Conversational Interaction in Japanese Language Learning

Xiling Liu¹  (✉),
Wei Liu² 

¹Tangshan College,
Tangshan, China

²Hangzhou DBAPPSecurity
Co., Ltd., Hangzhou, China

lxlsxy@163.com

ABSTRACT

Japanese language learning has long been constrained by acquisition difficulties arising from linguistic distance and by the limited availability of authentic opportunities for conversational interaction. Although mobile learning technologies offer considerable potential to alleviate these constraints, existing tools remain largely focused on vocabulary acquisition and grammar practice, while the dynamic progression of learner–artificial intelligence conversational interaction has received insufficient attention. In this study, a mobile Japanese conversational interaction system powered by a large language model was developed and integrated with automatic speech recognition and speech synthesis interfaces. Multi-topic conversational environments were constructed for learners at different proficiency levels. A four-week experiment involving 76 learners of Japanese was conducted, and interaction logs, Japanese-Language Proficiency Test–aligned mock assessments, and expert ratings were collected to characterize trajectories of conversation deepening and examine their associations with language learning outcomes. Pseudotime trajectory inference based on a minimum spanning tree demonstrated high internal consistency and external validity, and the composite conversation-depth score accounted for 43.8% of the variance in learner performance. Within individual conversational sessions, conversation depth increased significantly as the number of interactions turns increased. Intermediate learners exhibited greater mean pseudotime progression, more branch points, and faster rates of conversation deepening than beginner learners. These findings demonstrate that temporal modeling of interaction trajectories can be used to represent conversation deepening in generative artificial intelligence–supported mobile Japanese language learning as a quantifiable dynamic process. This approach provides a quantitative basis for investigating mobile conversation-based learning mechanisms and informing the design of adaptive conversational interactions.

KEYWORDS

generative artificial intelligence, Japanese language learning, mobile-assisted language learning, conversational interaction, second language acquisition

Liu, X., Liu, W. (2026). A Generative Artificial Intelligence-Based Mechanism for Mobile Conversational Interaction in Japanese Language Learning. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(20), pp. 78–92. <https://doi.org/10.3991/ijim.v20i20.63476>

Article submitted 2026-06-08. Revision uploaded 2026-08-28. Final acceptance 2026-09-12.

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

1 INTRODUCTION

The rapid development of generative artificial intelligence is fundamentally reshaping the modes of interaction available for mobile-assisted language learning [1]. Large language models, exemplified by the GPT series, have enabled mobile learners to engage in open-ended, natural, and real-time target-language conversation with artificial intelligence agents through their advanced capabilities in conversation generation and contextual understanding [2]. In parallel, the widespread adoption of mobile-assisted language learning has extended language learning beyond conventional classroom settings into fragmented learning opportunities embedded in everyday life [3], thereby allowing conversational practice to be undertaken anytime and anywhere with reduced anxiety and increased frequency [4]. The convergence of these technological advances and increasingly flexible learning contexts has facilitated the emergence of numerous generative artificial intelligence-supported mobile language learning applications, through which interaction data of unprecedented scale and ecological validity have become available for research on second language acquisition [5].

However, a substantial gap has emerged between the rapid evolution of technological tools and the explanatory capacity of existing theoretical frameworks. Three major limitations can be identified in the current literature. First, research has been dominated by outcome-oriented evaluation, whereas process-oriented analysis has remained comparatively underdeveloped [6]. Pre- and post-test designs have been widely used to examine the effects of generative artificial intelligence-based conversation systems on language learning outcomes [7], while questionnaire-based approaches have been employed to assess changes in learners' technology acceptance and self-efficacy [8]. Although such studies have established whether measurable learning effects occur, the dynamic processes through which these effects emerge have remained insufficiently understood. Fine-grained analyses of interaction logs have recently begun to receive attention [9]; however, most have remained limited to static content coding or frequency-based analyses, with little temporal modeling of the dynamic evolution of conversational interaction [10]. Second, the temporal and contextual characteristics of mobile interaction have not been adequately incorporated into existing analytical approaches. Mobile-assisted language learning is distinguished from desktop-based learning by fragmented learning contexts, intermittent interaction, and heterogeneous usage settings [11]. Conversational interaction on mobile devices inherently exhibits a temporal structure [12]: intervals occur between individual conversations, interaction turns unfold sequentially within each conversation, and longer-term developmental trajectories may emerge across conversations [13]. Nevertheless, individual conversation turns have generally been treated as independent observations, thereby overlooking temporal dependencies and developmental relationships across successive turns. Third, a computational framework for quantifying conversation deepening remains lacking. Although deep interaction has been extensively discussed within the field of second language acquisition, a system of quantitative indicators that can be computed in real time within mobile learning environments has yet to be established. Consequently, the current stage of conversation deepening cannot be dynamically identified by the system, and artificial intelligence response strategies cannot be readily adapted on this basis to facilitate deeper negotiation of meaning.

To address these limitations, an analytical framework based on temporal modeling of interaction trajectories is proposed and empirically validated for investigating the dynamics of mobile conversation learning. Specifically, the pseudotime trajectory inference paradigm originally developed in single-cell transcriptomics is adapted to the analysis of mobile human–artificial intelligence interaction. Multi-turn conversations between learners and artificial intelligence agents at different time points are represented as observations distributed along a latent conversation-deepening continuum, and trajectory inference is used to reconstruct the progression from shallow social exchanges to deeper negotiation of meaning. On this basis, a full-trajectory data acquisition and feature engineering framework for mobile conversational interaction is developed, incorporating more than 20 features across three dimensions: linguistic, interactional, and contextual. A multidimensional framework for quantifying conversation depth is further established, providing mobile conversational systems with a depth-awareness signal that can be computed in real time.

2 MOBILE CONVERSATIONAL INTERACTION SYSTEM DESIGN

A generative artificial intelligence-supported mobile application for Japanese conversation learning was developed using a three-tier architecture comprising a client, a server, and an artificial intelligence engine. The three components were decoupled to enable high-density interaction data to be collected for subsequent temporal modeling of interaction trajectories. The client was developed using the Flutter framework and was deployed on both iOS and Android platforms, with text- and speech-based input supported. Automatic speech recognition was performed locally on the mobile device using Faster-Whisper to reduce transmission latency and minimize privacy risks. Speech output was generated using Microsoft Edge text-to-speech with the Japanese ja-JP-Nanami voice. An asynchronous backend was implemented using FastAPI and was responsible for maintaining conversation states, orchestrating large language model requests, and persistently storing interaction logs. The artificial intelligence engine dynamically controlled the lexical difficulty and grammatical complexity of model-generated responses according to each learner's Japanese-Language Proficiency Test level. Five proficiency levels were accommodated through prompt templates combining few-shot exemplars with explicit constraint-based instructions. For each conversational interaction, a comprehensive set of metadata was automatically recorded, including timestamps and turn numbers, speaker identifiers, original and de-identified message text, input modality and speech duration, artificial intelligence response latency, device screen state and foreground/background transition indicators, network connection type, and conversation identifiers. These fields were specifically defined to support the linguistic, interactional, and contextual dimensions required for subsequent feature engineering, thereby preserving the information required to transform raw interaction logs into structured feature vectors. The overall framework for temporal modeling and depth-aware analysis of generative artificial intelligence-supported mobile Japanese conversational interaction is illustrated in Figure 1.

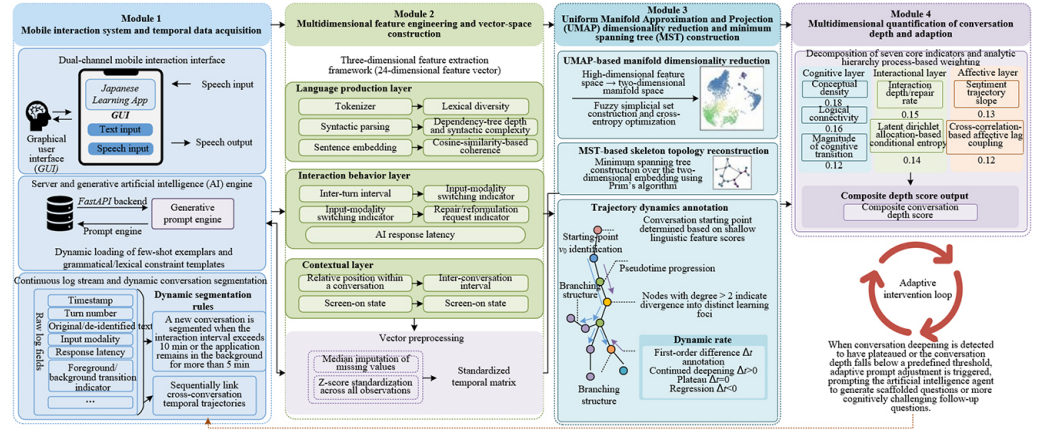


Fig. 1. Overall framework for temporal modeling and depth-aware analysis of generative artificial intelligence-supported mobile Japanese conversational interaction

3 INTERACTION-TRAJECTORY FEATURE ENGINEERING

The raw interaction logs were chronologically indexed by timestamp, with each individual message treated as the basic unit of observation. Before feature extraction, the continuous temporal logs were reorganized into semantically bounded conversation units through conversation segmentation. A conversation was considered to begin when the first message was sent by a learner or when the interval between the current and preceding messages exceeded 10 minutes. A conversation was considered to end when the learner explicitly exited the conversation, when no interaction occurred for more than 10 minutes, or when the application remained in the background for more than 5 minutes. These thresholds were selected with reference to established practices for distinguishing within-conversation continuity from separate conversations in fragmented mobile interaction contexts. Following segmentation, each conversation was represented as an ordered sequence of interaction turns. All conversations associated with each learner were then concatenated in chronological order to construct a complete interaction trajectory, thereby providing the data foundation for subsequent cross-conversation temporal analysis.

To transform the raw conversation text and interaction logs into numerical representations suitable for trajectory inference, a feature extraction framework was constructed across three dimensions: linguistic, interactional, and contextual. A 24-dimensional feature vector was ultimately generated for each observation. Linguistic features were extracted separately from the learner message and artificial intelligence response within each interaction turn and comprised four categories: lexical diversity, mean sentence length, syntactic complexity, and topic coherence. Lexical diversity was quantified using the type–token ratio, defined as the number of unique word types divided by the total number of word tokens. Japanese tokenization was performed using the Sudachi tokenizer. Mean sentence length was calculated directly as the number of characters per sentence. Syntactic complexity was quantified through dependency parsing and was calculated as follows:

$$D_t^{dep} = \frac{1}{M_u + M_a} \sum_{i=1}^{M_u + M_a} depth_i \quad (1)$$

where, M_u and M_a denote the numbers of bunsetsu units in the learner message and artificial intelligence response, respectively, and $depth_i$ denotes the depth of

the i -th bunsetsu unit in the dependency tree. Dependency parsing was performed using SuPar-UniDic, a Bidirectional Encoder Representations from Transformers (BERT)-based Japanese dependency parser. Topic coherence was quantified using the cosine similarity between the BERT-based sentence embeddings of the conversational content in adjacent turns:

$$\cos_sim(e_t, e_{t-1}) = \frac{e_t \cdot e_{t-1}}{\|e_t\| \|e_{t-1}\|} \quad (2)$$

Sentence representations were extracted using the pretrained Japanese BERT model `tohoku-nlp/bert-base-japanese`, with the final-layer hidden state corresponding to the [CLS] token used as the representation of the entire input sequence. This metric was used to characterize the temporal continuity of semantic content across successive conversation turns. Features within the interactional dimension were extracted to characterize learner engagement strategies and interaction-related behaviors, including the inter-turn interval, input-modality switching indicator, artificial intelligence response latency, and repair-request indicator. The inter-turn interval was calculated in seconds as $\Delta\tau_t = \tau_t - \tau_{t-1}$, where τ_t denotes the timestamp of the t -th interaction turn. The input-modality switching indicator I_t^{switch} was assigned a value of 1 when the input modality m_t of the current turn differed from that of the preceding turn m_{t-1} , and 0 otherwise, where $m_t \in \{text, voice\}$. The repair-request indicator, $r_t \in \{0, 1\}$, was determined through rule-based pattern matching and was assigned according to whether the learner message contained Japanese expressions requesting that the artificial intelligence agent rephrase, simplify, or clarify its preceding response. Features within the contextual dimension were extracted from the mobile usage context and included normalized within-conversation position, inter-conversation interval, and device-use state. These features were used to characterize temporal and contextual patterns of learner interaction across fragmented mobile learning contexts. The normalized within-conversation position was calculated as $p_t = t/N$, where t denotes the ordinal position of the current turn within a conversation and N denotes the total number of turns in that conversation. The inter-conversation interval was defined as the elapsed time ΔT_s between the start of the current conversation and the end of the preceding conversation; for the first conversation associated with each learner, this value was set to 0.

After all extracted features were concatenated, each interaction turn was represented by a complete feature vector, $x_t = [TTR_t, L_t, D_t^{dep}, \cos_sim(e_t, e_{t-1}), \Delta\tau_t, I_t^{switch}, \delta_t, r_t, p_t, \Delta T_s, screen_state_t, bg_switch_t, \dots]^T \in \mathbb{R}^{24}$, where δ_t denotes the artificial intelligence response latency, $screen_state_t \in \{0, 1\}$ is the screen state, and bg_switch_t is the number of foreground/background transitions within the conversation. Accordingly, the complete interaction trajectory of each learner was represented as an ordered sequence of feature vectors, $\{x_1, x_2, \dots, x_M\}$, where M denotes the total number of interaction turns recorded for that learner during the study period. Because substantial differences in scale were present across the extracted features, Z-score standardization was applied before trajectory inference:

$$x_t^{(k), norm} = (x_t^{(k)} - \mu_k) / \sigma_k \quad (3)$$

where, μ_k and σ_k denote, respectively, the mean and standard deviation of the k -th feature across all observations. Missing values, which accounted for less than 0.5% of the feature-level observations, were imputed using the median of the

corresponding feature. Standardization was subsequently performed to place all features on a comparable scale before trajectory inference. Through this feature-engineering framework, unstructured conversational interactions were transformed into structured temporal representations, thereby providing a computational basis for reconstructing trajectories of conversation deepening.

4 PSEUDOTIME INFERENCE MODEL FOR CONVERSATION-DEEPENING TRAJECTORIES

To characterize how Japanese conversational interaction between learners and generative artificial intelligence progresses from shallow social exchanges toward deeper negotiation of meaning in mobile learning environments, the pseudotime trajectory inference paradigm originally developed to reconstruct cellular differentiation trajectories in single-cell transcriptomics was adapted to mobile human–artificial intelligence interaction analysis. Under this framework, each interaction turn was represented as an observation in a high-dimensional feature space. The complete set of observations was defined as $X = \{x_1, x_2, \dots, x_N\} \in \mathbb{R}^{N \times d}$, where N denotes the total number of interaction turn across all learners and $d = 24$ denotes the dimensionality of the feature space. Although individual interaction turns occurred at different chronological times, each observation was assumed to occupy a particular position within a latent interaction-state space characterized by linguistic complexity and degree of interactional engagement. By ordering these observations along an inferred trajectory, a continuous representation of conversation deepening was reconstructed. Because distance-based relationships can become less discriminative in high-dimensional spaces, dimensionality reduction was performed before trajectory reconstruction.

The uniform manifold approximation and projection method was used to map the standardized high-dimensional feature vectors into a two-dimensional manifold representation, $Y \in \mathbb{R}^{N \times 2}$. Unlike principal component analysis and t -distributed stochastic neighbor embedding, the uniform manifold approximation and projection method can preserve both the local neighborhood structure and the global topological relationships of the data, making it well suited to conversational data characterized by both local temporal similarities and global evolutionary trends. The theoretical foundation of the method is based on the topological representation of fuzzy simplicial sets, with the core assumption that high-dimensional data are distributed approximately uniformly in the vicinity of an underlying low-dimensional manifold. Dimensionality reduction was performed in two principal stages. First, a fuzzy topological representation was constructed in the high-dimensional feature space. For each observation x_i , the directed fuzzy membership strength between x_i and a neighboring observation x_j was defined as:

$$v_{ji} = \exp(-(d(x_i, x_j) - \rho_i)/\sigma_i) \quad (4)$$

where, $d(x_i, x_j)$ denotes the Euclidean distance, ρ_i is a local connectivity parameter determined by the distance from x_i to its nearest neighbor, and σ_i is an adaptive local scaling parameter determined through $\sum_{j=1}^k v_{ji} = \log_2(k)$, with $k = 15$ as the predefined number of nearest neighbors. After being subsequently symmetrized, $v_{ij} = v_{ji} + v_{ij} - v_{ji} \cdot v_{ij}$ were obtained.

In the low-dimensional embedding space, the fuzzy membership strength between y_i and y_j was defined as:

$$\omega_{ij} = \left(1 + a \cdot \|y_i - y_j\|_2^{2b}\right)^{-1} \quad (5)$$

where, $a \approx 1.93$ and $b \approx 0.79$ are approximation parameters. The low-dimensional embedding was then obtained by minimizing the cross-entropy loss:

$$L_{UMAP} = \sum_{i \neq j} \left[v_{ij} \log(v_{ij}/\omega_{ij}) + (1 - v_{ij}) \log((1 - v_{ij})/(1 - \omega_{ij})) \right] \quad (6)$$

Optimization was performed using stochastic gradient descent for 500 epochs.

Following dimensionality reduction, a minimum spanning tree was constructed in the two-dimensional manifold space to approximate the backbone of the conversation-deepening trajectory. The minimum spanning tree was defined as an acyclic connected subgraph spanning all observations while minimizing the total edge weight. Because no explicit parametric form of the trajectory was imposed, the resulting structure was determined entirely by the intrinsic organization of the data. Edge weights were defined using the Euclidean distance $w(i, j) = \|y_i - y_j\|_2$, and Prim's algorithm was used. The starting point of the trajectory was identified as the observation with the lowest linguistic feature score across all conversational turns, representing the initial state of conversation deepening. Lexical diversity, mean sentence length, and syntactic complexity were first normalized to the interval $[0, 1]$, and a shallow-interaction score was calculated for each interaction turn:

$$S_t^{shallow} = -(TTR_t^{norm} + L_t^{norm} + D_t^{dep, norm})/3 \quad (7)$$

The lowest 5% of observations were selected to form the candidate starting-point set S , and the mean of their coordinates in the reduced-dimensional space was calculated as the reference center $y_{start} = |S|^{-1} \sum_{y_t \in S} y_t$. The actual observed node closest to this reference center was then identified as $y_{start'}$, from which traversal along the minimum spanning tree was initiated. The shortest-path distance was defined as the pseudotime value of the corresponding conversational turn:

$$\psi_t = \min_{P \in \mathcal{P}(start, y_t)} \sum_{(i, j) \in P} \|y_i - y_j\|_2 \quad (8)$$

where, $\mathcal{P}(start, y_t)$ denotes the set of all paths from the starting point to y_t . The tree structure guarantees the uniqueness of the path between any two nodes, allowing the path distances to be computed in linear time using depth-first search. The resulting pseudotime values were bounded by $[0, \Psi_{max}]$, where $\Psi_{max} = \max_t \psi_t$. After normalization, $\psi_t^{norm} = \psi_t / \Psi_{max} \in [0, 1]$, with $\psi_t^{norm} = 0$ corresponding to the trajectory starting point and $\psi_t^{norm} = 1$ corresponding to the trajectory endpoint.

The inferred conversation-deepening trajectory was not assumed to follow a single linear pathway, as interaction patterns may diverge after comparable levels of progression have been reached. To identify potential branching structures, nodes with degree greater than 2 in the minimum spanning tree were designated as branch points, and the subtrees extending from these nodes were treated as candidate divergent trajectories. Branch specificity was quantified using the Mahalanobis

distance between the mean feature vector of observations within a given branch and that of observations occupying the corresponding region of the principal trajectory. A higher score indicates that the branch exhibits more distinctive linguistic characteristics. In addition to static trajectory positioning, the rate of conversation deepening was characterized from a dynamic perspective. The first-order difference in pseudotime, $\Delta\psi_t = \psi_t - \psi_{t-1}$, represents the increment in conversation depth at each turn and was set to zero for the first turn of each conversation. $\Delta\psi_t > 0$ indicates continued deepening, $\Delta\psi_t \approx 0$ indicates a plateau, and $\Delta\psi_t < 0$ indicates regression in conversation depth. The cumulative deepening rate, $R_{cum} = \frac{1}{M} \sum_{t=1}^M \Delta\psi_t$, represents the average pseudotime increment per conversational turn, thereby integrating the extent and efficiency of conversation deepening within a unified framework and providing a quantitative basis for comparing differences in the efficiency of deepening trajectories across learners.

5 MULTIDIMENSIONAL FRAMEWORK FOR QUANTIFYING CONVERSATION DEPTH

Pseudotime trajectory inference characterizes the relative position and progression of conversation along an inferred conversation-deepening continuum but does not, by itself, establish the extent to which a given interaction constitutes substantively deep conversation with potential relevance to language learning. To address this limitation, a set of real-time computable indicators was developed to quantify conversation depth across three dimensions: cognitive, interactional, and affective. The cognitive dimension comprised three indicators: conceptual density, logical connectivity, and semantic transition magnitude. Conceptual density was used to quantify the amount of distinct conceptual information expressed per unit of conversation length and was defined as $\rho_{cog} = V_{unique}/N_{total}$, where V_{unique} denotes the number of unique concepts identified from noun and verb phrases extracted through Japanese dependency parsing, and N_{total} denotes the total number of word tokens. Logical connectivity, $\lambda = (M-1)^{-1} \sum_{t=2}^M \cos_sim(e_t, e_{t-1})$, was quantified as the mean cosine similarity between the BERT-based sentence embeddings of adjacent conversational turns, with $\lambda \in [-1, 1]$, thereby capturing the temporal continuity of conversational semantics. Higher values indicate stronger logical coherence, making this metric particularly useful for distinguishing meaningful, in-depth conversations from shallow conversations characterized by abrupt or random topic shifts. The magnitude of cognitive transition, $\kappa_t = \|e_t - e_{t-1}\|_2 / \max_t \|e_t - e_{t-1}\|_2$, was used to capture moments of conceptual breakthrough, with the denominator representing the maximum semantic transition magnitude across all samples. The resulting κ_t was normalized to the range $[0, 1]$, with values approaching 1 indicating a pronounced qualitative shift in cognition.

The interactional dimension comprised two indicators reflecting learner-initiated engagement and topic-management behavior: repair engagement and topic persistence. Repair engagement, $\rho_{inter} = T^{-1} \sum_{k=1}^T r_k$, was defined as the frequency with which a learner actively requested the artificial intelligence agent to revise or refine its responses within the current conversation, where T denotes the total number of conversational turns and $r_k \in \{0, 1\}$ denotes the repair-request indicator. This metric reflects the learner's level of metacognitive monitoring. Topic persistence,

$\sigma_t = 1 - H(\text{topic}_t | \text{topic}_{t-1}) / H_{\max}$, quantifies the extent to which a topic is consistently maintained across consecutive conversational turns rather than being subject to frequent shifts. The conditional entropy $H(\text{topic}_t | \text{topic}_{t-1})$ was estimated using a latent Dirichlet allocation model with $K = 20$ topics, as follows:

$$H(\text{topic}_t | \text{topic}_{t-1}) = - \sum_{i,j} P(\text{topic}_t = i, \text{topic}_{t-1} = j) \log P(\text{topic}_t = i | \text{topic}_{t-1} = j) \quad (9)$$

Higher values of $H_{\max} = \log_2(K)$, $\sigma_t \in [0, 1]$ indicated greater topic persistence across successive interaction turns. The affective dimension was characterized in terms of affective dynamics and the coupling between the learner and the artificial intelligence agent. The affective-valence slope was used to quantify the direction and magnitude of change in affective valence over the course of a conversation and was estimated as $\beta_{\text{aff}} = \sum_{t=1}^N (t - \bar{t})(s_t - \bar{s}) / \sum_{t=1}^N (t - \bar{t})^2$, where $s_t \in [-1, 1]$ denotes the affective-valence score generated by a Japanese-language sentiment classification model, $\bar{t} = (N + 1) / 2$ denotes the mean turn index, and $\bar{s} = N^{-1} \sum_{t=1}^N s_t$ denotes the mean affective-valence score within the conversation. $\beta_{\text{aff}} > 0$ indicates an increasingly positive affective trajectory over successive interaction turns. Affective coupling lag was used to characterize the temporal association between changes in learner and artificial intelligence affective expressions. The cross-correlation function was calculated as follows:

$$R_{TS}(L) = |T_L|^{-1} \sum_{t \in T_L} s_t^T \cdot s_{t+L}^A \quad (10)$$

where, s_t^T and s_t^A denote the affective-valence sequences derived from learner messages and artificial intelligence responses, respectively; $T_L = \{t | 1 \leq t + L \leq N\}$ denotes the set of valid overlapping indices; and $L \in [-3, 3]$ denotes the lag measured in interaction turns. The sign of affective coupling lag, $L^* = \arg \max_{L \in [-3, 3]} |R_{TS}(L)|$, indicates whether the learner or the artificial intelligence agent leads the affective dynamics, whereas its absolute magnitude reflects the strength of affective coupling.

The seven component indicators were integrated into a composite conversation depth score, $D_t = w_{\rho} \rho_{\text{cog}} + w_{\lambda} \lambda + w_{\kappa} \kappa_t + w_{\rho} \rho_{\text{inter}} + w_s \sigma_t + w_a \beta_{\text{aff}} + w_c |L^*|$, using a weighted summation. Indicator weights were determined using the analytic hierarchy process based on judgments provided by three experts in Japanese language education and two researchers specializing in human-computer interaction. The relative importance of the seven indicators was evaluated through pairwise comparisons using a 1–9 rating scale. The consistency ratio of the resulting judgment matrix was below 0.10, indicating an acceptable level of consistency in the expert judgments. The final weights assigned to conceptual density, logical connectivity, semantic transition magnitude, repair engagement, topic persistence, affective-valence slope, and affective coupling were 0.18, 0.16, 0.12, 0.15, 0.14, 0.13, and 0.12, respectively. The resulting composite conversation depth score, D_t , was designed to provide a computational signal that could support real-time depth-aware adaptation in the mobile conversation system. When the score for a given conversational turn fell below a predefined threshold, the system could prompt the learner to engage in more in-depth expression or pose more cognitively challenging questions, thereby translating the theoretical framework into an active intervention mechanism within the system.

6 EXPERIMENTAL DESIGN AND RESULTS

6.1 Internal validation of trajectory inference

The reliability of the minimum spanning tree-based pseudotime trajectory inference procedure for reconstructing conversation-deepening trajectories was evaluated using repeated random holdout validation. In each iteration, 20% of the interaction turns were randomly excluded from trajectory inference, and the minimum spanning tree-based trajectory was reconstructed using the remaining 80% of the observations. The excluded observations were subsequently projected onto the reconstructed trajectory to obtain predicted pseudotime values, $\hat{\psi}_t$. Spearman's rank correlation coefficient was then calculated between the predicted pseudotime values and the corresponding actual pseudotime values, ψ_t . To reduce dependence on a particular random partition, the holdout procedure was repeated 10 times.

The results of the 10 repeated random holdout validations are presented in Table 1. Spearman's rank correlation coefficient ρ_s averaged 0.880 (standard deviation = 0.007) and ranged from 0.868 to 0.891 across the 10 repetitions, with $p < 0.001$ in every iteration. These results demonstrate a strong positive correlation between the predicted pseudotime and the ground-truth pseudotime for the held-out rounds, indicating that the minimum spanning tree-based pseudotime trajectory inference exhibits high internal consistency and can robustly reconstruct the trajectory of dialogue progression. To further assess external validity, three Japanese language education experts, each with more than five years of teaching experience, were invited to independently rate the depth of 100 randomly selected dialogue segments using a five-point scale. The mean expert ratings were strongly correlated with the corresponding pseudotime values, ψ_t^{norm} , with a Pearson correlation coefficient of 0.782 ($p < 0.001$). This finding indicates a high degree of agreement between the pseudotime ordering inferred from the trajectory analysis and the depth assessments provided by human experts.

Table 1. Results of repeated random holdout validation across 10 repetitions

Measure	1	2	3	4	5	6	7	8	9	10	Mean	Standard Deviation
Held-out interaction turns	7968	7968	7968	7968	7968	7968	7968	7968	7968	7968	7968	–
Retained interaction turns	31874	31874	31874	31874	31874	31874	31874	31874	31874	31874	31874	–
ρ_s	0.872	0.884	0.879	0.891	0.868	0.885	0.877	0.890	0.882	0.876	0.880	0.007
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	–

6.2 Predictive validity of the conversation depth score

The predictive validity of the composite conversation depth score, D_c , was evaluated by examining its association with language learning outcomes. Before and after the intervention, learners completed Japanese-Language Proficiency Test-based mock assessments corresponding to their proficiency levels, ranging from N5 to N2. Learning gain was defined as the change in assessment score from pretest to posttest, $\Delta\text{Score} = \text{Post} - \text{Pre}$, with each assessment scored on a 100-point scale. For each learner, interaction turns were aggregated by conversation, and the mean

composite conversation depth score was calculated separately for each conversation. A multiple linear regression model, $\Delta Score = \alpha + \sum_{k=1}^K \gamma_k \bar{D}_k + \epsilon$, was then constructed, where K represents the total number of sessions for each learner, with session order included as a control variable.

The parameter estimates of the multiple linear regression model are presented in Table 2. The model demonstrated an overall goodness of fit of $R^2 = 0.438$, with an adjusted R^2 of 0.412 ($F(8, 67) = 6.532$, $p < 0.001$). The $\bar{D}_k$ coefficients for the first six conversations were all significantly positive ($p < 0.05$), with Conversation 4 exhibiting the largest coefficient ($\gamma_4 = 7.245$, $p = 0.004$). Specifically, a 0.1-point increase in conversation depth during Conversation 4 was associated with an expected increase of approximately 0.72 points in the learner's post-test score. From Conversation 7 onward, the coefficients gradually became statistically nonsignificant. This pattern may reflect a learning plateau effect, whereby the marginal contribution of conversation depth to learning performance diminishes once learners reach a certain level of proficiency. Overall, the composite conversation depth score explained 43.8% of the variance in learner performance, indicating moderate-to-strong predictive validity.

Table 2. Multiple linear regression results for predicting $\Delta score$

Predictor	Coefficient α	$\bar{D}_1$ (Conversation 1)	$\bar{D}_2$ (Conversation 2)	$\bar{D}_3$ (Conversation 3)	$\bar{D}_4$ (Conversation 4)
Coefficient γ	2.341	4.872	5.234	6.108	7.245
Standard error	1.872	2.103	2.051	2.214	2.398
t value	1.251	2.317	2.552	2.759	3.021
p value	0.215	0.024	0.013	0.008	0.004
95% confidence interval	[-1.398, 6.080]	[0.673, 9.071]	[1.139, 9.329]	[1.688, 10.528]	[2.457, 12.033]
Predictor	Coefficient α	$\bar{D}_5$ (Conversation 5)	$\bar{D}_6$ (Conversation 6)	$\bar{D}_7$ (Conversation 7)	$\bar{D}_8$ (Conversation 8+)
Coefficient γ	2.341	6.893	5.671	4.298	3.124
Standard error	1.872	2.287	2.356	2.442	2.561
t value	1.251	3.014	2.407	1.760	1.220
p value	0.215	0.004	0.019	0.083	0.227
95% confidence interval	[-1.398, 6.080]	[2.326, 11.460]	[0.967, 10.375]	[-0.578, 9.174]	[-1.989, 8.237]

6.3 Dynamic changes in conversation depth across turns within conversations

The within-conversation dynamics of conversation depth were examined to determine whether the composite conversation depth score increased systematically across successive interaction turns. A multilevel model was fitted to distinguish between conversation-level and turn-level variation. The dependent variable was the composite conversation depth score for an individual interaction turn, D_{ij} (turn i within conversation j). The predictors included turn number, $Turn_{ij}$, a within-conversation position variable, $IsWithin_{ij}$, and their interaction:

$$D_{ij} = \beta_0 + \beta_1 \cdot Turn_{ij} + \beta_2 \cdot IsWithin_{ij} + \beta_3 \cdot IsWithin_{ij} \times Turn_{ij} + u_{0j} + \epsilon_{ij}$$

where, $u_{0j} \sim N(0, \sigma_u^2)$ denotes the conversation-level random intercept, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ denotes the turn-level residual, and β_3 represents the coefficient of primary

interest. The parameter estimates from the multilevel model are presented in Table 3. A significant positive interaction effect, $IsWithin \times Turn$, was observed ($\beta_3 = 0.021$, $p < 0.001$), indicating that the composite depth score increased by an average of 0.021 points with each additional turn within the same conversation. For a representative 10-turn conversation, this estimate corresponds to an expected increase of approximately 0.189 points from the first to the tenth turn, corresponding to a progression from shallow dialogue ($D \approx 0.30$) to moderately deep dialogue ($D \approx 0.49$). The main effect of turn order was also significantly positive ($\beta_1 = 0.008$, $p = 0.008$), indicating a gradual but significant increase in conversation depth with the accumulation of overall usage experience, even across conversations. The intraclass correlation coefficient was 0.300, indicating that between-conversation differences accounted for 30.0% of the total variance in depth scores, whereas the remaining 70.0% was attributable to within-conversation variation. These findings support the necessity of analyzing the dynamic changes in conversation depth at the turn level. Figure 2 shows the parameter estimates of the multilevel model with random effects.

Table 3. Fixed-effects estimates from the multilevel model

Fixed Effect	Estimate	Standard Error	Degrees of Freedom	t Value	p Value	95% Confidence Interval
Intercept β_0	0.287	0.019	8419	15.105	<0.001	[0.250, 0.324]
Turn β_1	0.008	0.003	39832	2.667	0.008	[0.002, 0.014]
IsWithin β_2	0.052	0.021	8419	2.476	0.013	[0.011, 0.093]
IsWithin $\times$ Turn β_3	0.021	0.004	39832	5.250	<0.001	[0.013, 0.029]

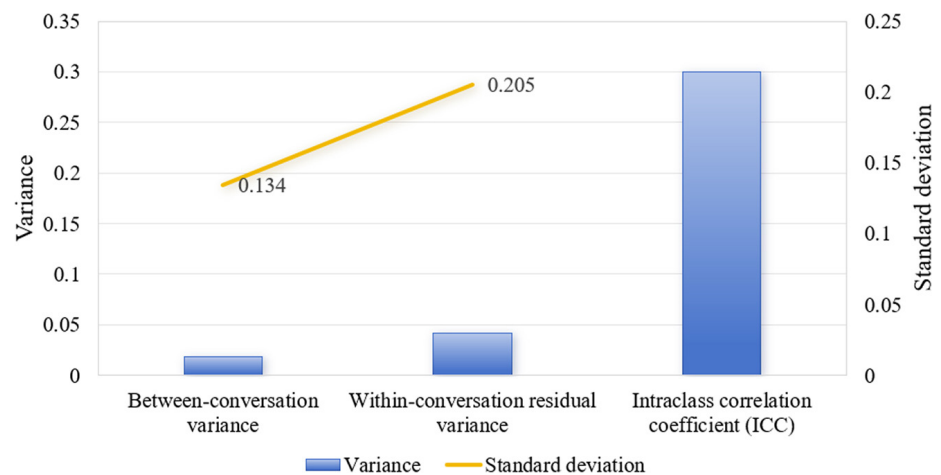


Fig. 2. Parameter estimates of the multilevel model with random effects

6.4 Comparison of conversation trajectories across proficiency levels

Differences in conversation-deepening trajectories were examined between beginner-level learners (N5–N4; $n_1 = 42$; pretest mean = 48.3, standard deviation = 11.2) and intermediate-level learners (N3–N2; $n_2 = 34$; pretest mean = 71.6, standard deviation = 9.8). Dynamic time warping was used to quantify differences between the mean trajectories of the two proficiency groups. Two-dimensional sequences defined jointly by pseudotime and conversation depth were aligned by minimizing

the cumulative distance between corresponding trajectory points. Three additional trajectory characteristics were compared between groups: mean pseudotime progression ($\bar{\psi}$), number of branch points, and mean pseudotime progression rate (R_{cum}).

The between-group comparisons are presented in Table 4. Significant differences between beginner- and intermediate-level learners were observed across all three trajectory measures ($p < 0.001$), with large standardized effect sizes (Cohen's $d > 0.8$). Mean pseudotime progression was significantly greater in the intermediate group ($\bar{\psi} = 0.603$) than in the beginner group ($\bar{\psi} = 0.412$, $t(74) = 9.384$, $p < 0.001$, Cohen's $d = 2.16$). This finding indicates that the conversations of intermediate-level learners were generally positioned further along the conversation-deepening axis. The mean number of branch points per learner was 3.4 in the intermediate group, compared with 1.8 in the beginner group ($t(74) = 6.553$, $p < 0.001$, Cohen's $d = 1.51$), indicating that intermediate-level learners explored a greater diversity of conversation-deepening trajectories. The significant difference in the conversation deepening rate, R_{cum} (1.42×10^{-3} vs. 0.87×10^{-3} , $t(74) = 6.940$, $p < 0.001$, Cohen's $d = 1.60$) further indicates that intermediate-level learners achieved significantly greater gains in conversation depth per turn. The dynamic time warping distance between the mean trajectories of the two groups was 0.347 in the normalized space, indicating a moderate difference in trajectory shape. Specifically, the trajectory of the beginner group exhibited a relatively smooth, monotonic increase, whereas that of the intermediate group showed multiple stepwise increases and branching structures in the middle-to-late portion of the pseudotime axis ($\psi \approx 0.5 - 0.8$). These patterns corresponded to breakthrough points associated with discussions of grammatical refinement and the practice of complex sentence structures.

Table 4. Comparison of conversation-trajectory measures between beginner- and intermediate-level learners

Measure	Beginner Group (n = 42)	Intermediate Group (n = 34)	Test Statistic	p Value	Effect Size (Cohen's d)
Mean pseudotime progression $\bar{\psi}$	0.412 (0.087)	0.603 (0.092)	$t(74) = 9.384$	<0.001	2.16
Number of branch points per learner	1.8 (0.9)	3.4 (1.2)	$t(74) = 6.553$	<0.001	1.51
Deepening rate $R_{cum} (\times 10^{-3})$	0.87 (0.31)	1.42 (0.38)	$t(74) = 6.940$	<0.001	1.60
Mean composite conversation depth score $\bar{D}$	0.394 (0.079)	0.557 (0.085)	$t(74) = 8.703$	<0.001	2.00
Mean number of interaction turns per conversation	8.7 (3.2)	12.4 (4.1)	$t(74) = 4.369$	<0.001	1.01

7 CONCLUSION

To address the limited understanding of interaction dynamics in generative artificial intelligence-supported mobile Japanese conversation learning, an analytical framework based on temporal modeling of interaction trajectories was developed, integrating data acquisition, feature engineering, trajectory inference, and depth quantification within a unified computational framework. By adapting pseudotime trajectory inference from single-cell transcriptomics to mobile human-artificial intelligence interaction, conversation deepening was represented as a

visualized, continuous, computationally tractable trajectory rather than as a series of isolated interaction outcomes. Experimental validation demonstrated that the proposed approach exhibited high internal consistency and external validity. The composite conversation depth score explained 43.8% of the variance in learner performance, and conversation depth increased significantly across successive turns within the same conversation. The findings further revealed that intermediate-level learners exhibited deeper, more diverse, and more rapidly progressing conversation trajectories than beginner-level learners.

These findings make contributions at theoretical, technological, and practical levels. Theoretically, the proposed framework provides computationally operationalized definitions and empirical support for the interaction hypothesis and the output hypothesis. Technologically, it provides mobile conversational systems with a quantitative tool for real-time assessment of interaction depth. Practically, the findings offer guidance for the design of dynamically adaptive generative artificial intelligence-assisted language learning systems. Several limitations should nevertheless be acknowledged, including the relatively limited sample size, the four-week experimental duration, the exclusive focus on Japanese, and the use of a single general-purpose large language model. Future research could incorporate reinforcement learning to develop strategies for actively guiding conversation deepening, extend the framework to multimodal interaction data to capture richer cognitive signals, and evaluate the proposed approach across additional languages and mobile interaction contexts to assess its generalizability.

8 REFERENCES

- [1] A. D. Samala, X. Zhai, K. Aoki, L. Bojic, and S. Zikic, "An in-depth review of ChatGPT's pros and cons for learning and teaching in education," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 2, pp. 96–117, 2024. <https://doi.org/10.3991/ijim.v18i02.46509>
- [2] M. Jamshed, A. S. M. M. Ahmed, M. Sarfaraj, and W. U. Warda, "The impact of ChatGPT on English language learners' writing skills: An assessment of AI feedback on mobile," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 19, pp. 18–36, 2024. <https://doi.org/10.3991/ijim.v18i19.50361>
- [3] R. Shadiev, T. Liu, and W. Hwang, "Review of research on mobile-assisted language learning in familiar, authentic environments," *British Journal of Educational Technology*, vol. 51, no. 3, pp. 709–720, 2019. <https://doi.org/10.1111/bjet.12839>
- [4] X. M. Li, M. Liu, and C. Zhang, "Technological impact on language anxiety dynamic," *Computers & Education*, vol. 150, Art. no. 103839, 2020. <https://doi.org/10.1016/j.compedu.2020.103839>
- [5] L. Guan, E. Y. Zhang, and M. M. Gu, "Examining generative AI-mediated informal digital learning of English practices with social cognitive theory: A mixed-methods study," *ReCALL*, vol. 37, no. 3, pp. 315–331, 2024. <https://doi.org/10.1017/S0958344024000259>
- [6] R. Wu and Z. Yu, "Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 10–33, 2023. <https://doi.org/10.1111/bjet.13334>
- [7] M. Li, Y. Wang, and X. Yang, "Can generative AI chatbots promote second language acquisition? A meta-analysis," *Journal of Computer Assisted Learning*, vol. 41, no. 4, Art. no. 70060, 2025. <https://doi.org/10.1111/jcal.70060>

- [8] F. Huang, Y. Wang, and H. Zhang, “Modelling generative AI acceptance, perceived teachers’ enthusiasm and self-efficacy to English as a foreign language learners’ well-being in the digital era,” *European Journal of Education*, vol. 59, no. 4, Art. no. e12770, 2024. <https://doi.org/10.1111/ejed.12770>
- [9] C. J. Lin and H. Mubarak, “Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom,” *Educational Technology & Society*, vol. 24, no. 4, pp. 16–35, 2021. <https://www.jstor.org/stable/48629242>
- [10] F. Ş. Koç and P. Savaş, “The use of artificially intelligent chatbots in English language learning: A systematic meta-synthesis study of articles published between 2010 and 2024,” *ReCALL*, vol. 37, no. 1, pp. 4–21, 2024. <https://doi.org/10.1017/S0958344024000168>
- [11] S. Figueiredo, “The effect of mobile-assisted learning in real language attainment: A systematic review,” *Journal of Computer Assisted Learning*, vol. 39, no. 4, pp. 1083–1102, 2023. <https://doi.org/10.1111/jcal.12811>
- [12] C. L. Nagle, “A design framework for longitudinal individual difference research: Conceptual, methodological, and analytical considerations,” *Research Methods in Applied Linguistics*, vol. 2, no. 1, Art. no. 100033, 2022. <https://doi.org/10.1016/j.rmal.2022.100033>
- [13] M. B. Paradowski, N. Whitby, M. Czuba, and P. Bródka, “Peer interaction dynamics and second language learning trajectories during study abroad: A longitudinal investigation using dynamic computational social network analysis,” *Language Learning*, vol. 74, no. S2, pp. 58–115, 2024. <https://doi.org/10.1111/lang.12681>

9 AUTHORS

Xiling Liu received her Master’s degree from Northeast Normal University in 2009 and her Ph.D. degree from Beijing Foreign Studies University in 2021. She is currently a faculty member at the School of Foreign Languages, Tangshan College. Her research interests include foreign linguistics and foreign language teaching (E-mail: lxltssxy@163.com).

Wei Liu studied at the China Institute of Defense Science and Technology and received his Bachelor’s degree in 2006. His research interests lie in Computer Science and Technology (E-mail: lwhzah@163.com).