AURA: A Reinforcement Learning Framework for AI-Driven Adaptive Conversational Surveys

Jinwen Tang, Yi Shang

University of Missouri, Department of Electrical Engineering and Computer Science, Columbia, MO, USA

Abstract

Conventional online surveys provide limited personalization, often resulting in low engagement and superficial responses. Although AI survey chatbots improve convenience, most are still reactive: they rely on fixed dialogue trees or static prompt templates and therefore cannot adapt within a session to fit individual users, which leads to generic follow-ups and weak response quality. We address these limitations with **AURA** (*Adaptive Understanding through Reinforcement Learning for Assessment*), a reinforcement learning framework for AI-driven adaptive conversational surveys. AURA quantifies response quality using a four-dimensional LSDE metric (Length, Self-disclosure, Emotion, and Specificity) and selects follow-up question types via an ϵ-greedy policy that updates the expected quality gain (expected value) within each session. Initialized with priors extracted from 96 prior campus-climate conversations (467 total chatbot-user exchanges), the system balances exploration and exploitation across 10-15 dialogue exchanges, dynamically adapting to individual participants in real time. In controlled evaluations, AURA achieved a +0.12 mean gain in response quality and a statistically significant improvement over non-adaptive baselines ($p = 0.044$, $d = 0.66$), driven by a 63% reduction in specification prompts and a 10× increase in validation behavior. These results demonstrate that reinforcement learning can give survey chatbots improved adaptivity, transforming static questionnaires into interactive, self-improving assessment systems.

---

1. Introduction

Campus-climate surveys play a crucial role in understanding how students, faculty, and staff experience everyday life at universities. By gathering feedback on issues such as academic support, discrimination, resource accessibility, and sense of belonging, these surveys enable institutions to identify structural barriers and design interventions that promote equity and inclusion (Vogel et al., 2008; PASSHE, 2022; Bordieri et al., 2024). When administered on a regular cadence, they also provide a longitudinal view of institutional progress, ensuring that improvement efforts remain evidence based and responsive to community needs.

However, the effectiveness of traditional survey instruments is constrained by their static format. Respondents often experience low engagement, survey fatigue, or limited opportunity for clarification, which can lead to superficial or incomplete data (Dillman et al., 2014; Galesic and Bosnjak, 2009; Sahlqvist et al., 2011). The impersonal, time consuming nature of conventional questionnaires further discourages participation and can undermine the reliability of collected information (Porter and Whitcomb, 2005). As a result, administrators may fail to capture the nuanced attitudes, emotions, and contextual factors that shape lived experiences on campus.

Recent advances in generative artificial intelligence (GenAI) and large language models (LLMs) have created opportunities to transform these processes. Conversational AI systems can conduct interactive, context-aware dialogues that encourage deeper reflection (Xiao et al., 2020; Tanwar et al., 2024). Early deployments in campus settings demonstrate that chatbot-based surveys can yield richer qualitative feedback, improved user satisfaction, and reduced survey fatigue compared to traditional web forms (Abbas et al., 2021; Kim et al., 2019; Zarouali et al., 2024).

Despite these advances, most existing AI-driven survey chatbots remain reactive rather than truly adaptive. Many still depend on rigid dialogue trees or static prompt templates that fails to evolve with user behavior during a session (Thorat and Jadhav, 2020; Maeng and Lee, 2021; Chen et al., 2024). Earlier frameworks such as TigerGPT (Tang et al., 2025), which combined

2

role-based onboarding with empathetic conversational design, also followed pre-scripted flows. These systems showed that generative dialogue can boost engagement, yet they lacked a way to learn from user feedback and adjust questioning strategies as the conversation unfolded.

This limitation narrows the potential of GenAI-driven surveys. Human interviewers naturally shift their style based on subtle cues — brevity, emotional tone, or how specific a response is — but static prompts cannot reproduce this kind of flexible reasoning. Work in dialogue management has shown that reinforcement learning can guide sequential decision making and improve both adaptability and personalization (Li et al., 2016; Su et al., 2021; Sun et al., 2023). More recent studies extend this approach to large language models, refining their decision policies through offline goal-conditioned optimization (Hong et al., 2025). Even so, reinforcement learning has rarely been applied to survey or assessment contexts, where adaptive behavior could meaningfully raise data quality and respondent trust.

A persistent challenge is *within-session personalization*: how can a chatbot learn, inside a single exchange, which questioning strategy best fits each individual? Some respondents engage deeply when asked for concrete examples, while others prefer brief factual prompts. Emotional reflection may help some participants but discourage others. Current systems typically rely on fixed sequences or simple rule-based branching and therefore fail to recognize these individual patterns (Xiao et al., 2020; Belhaj et al., 2021). In campus-climate datasets, for instance, 64.2% of responses show low engagement (composite quality scores below 0.4), suggesting that static question logic struggles to sustain participation across diverse users.

This gap highlights the missing connection between conversational design and computational learning. Human-computer-interaction frameworks emphasize empathy, personalization, and user-controlled flow (Kim et al., 2008; Kocaballi et al., 2019), whereas reinforcement learning formalizes decision making through reward feedback (Sutton et al., 1998). Bringing these perspectives together would allow an AI survey agent to align conversational empathy with measurable engagement outcomes — transforming reactive dialogue into an adaptive, continuously improving interaction process that better reflects human interviewing practice.

To address this gap, we present **AURA** (*Adaptive Understanding through Reinforcement Learning for Assessment*), a reinforcement-learning–enhanced survey framework that learns optimal questioning strategies within individual conversations. AURA extends the conversational design principles of our

pilot system, TigerGPT (Tang et al., 2025), with three key innovations:

Real-time engagement measurement. AURA quantifies user engagement through a four-dimensional quality metric combining *Length*, *Self-disclosure*, *Emotion*, and *Specificity* (LSDE). This composite score is computed after each user response and mapped to discrete engagement states that guide subsequent action selection.

Data-driven action selection. Rather than following fixed question sequences, AURA maintains an expected-value (EV) table derived from 96 prior campus-climate conversations (467 total exchanges). Given the current engagement state, the system selects among five question types—specification, elaboration, topic probe, validation, and continuation—to maximize predicted quality improvement.

Within-session learning. Through an ϵ-greedy policy ($\epsilon = 0.30$, $\alpha = 0.30$), AURA balances exploration of new strategies with exploitation of successful patterns. The EV estimates are updated at each exchange, allowing the policy to adapt to individual users within approximately 10–15 conversational exchanges.

This design replaces reactive conversational heuristics with a self-improving policy that learns as dialogue unfolds. By initializing from empirical data but updating in real time, AURA overcomes the cold-start problem and remains responsive to individual differences in engagement style.

This study makes four contributions to adaptive survey research and conversational AI:

- **Reinforcement Learning Framework for Adaptive Surveys:** AURA introduces an RL-enhanced framework that learns optimal question-selection strategies within individual conversations, bridging conversational-design theory with computational learning.

- **Four-Dimensional Engagement Metric:** A validated LSDE scoring framework that quantifies real-time response quality across length, self-disclosure, emotion, and specificity, enabling automatic measurement of engagement during dialogue.

- **Prior Pattern Initialization:** A data-driven method to extract expected-value priors from 96 authentic campus-climate conversations (467 exchanges), addressing the cold-start problem while supporting rapid within-session adaptation.

4

- **Empirical Validation:** Quantitative and behavioral analyses demonstrate a $+0.12$ mean gain in response quality and statistically significant improvements ($p = 0.044$, $d = 0.66$) under the optimal configuration ($\epsilon = 0.30$, $\alpha = 0.30$), confirming that within-session reinforcement learning enhances conversational-survey engagement.

Broader implications. AURA demonstrates a generalizable approach to adaptive conversational systems for sensitive data collection. Its ability to learn from limited prior data (96 conversations) while adapting within short interactions (10–15 exchanges) suggests applications in mental-health screening, patient-intake interviews, qualitative research, and workplace-climate surveys — settings where traditional static forms yield poor engagement but human-interviewer adaptivity is cost-prohibitive. By formalizing the connection between conversational-quality metrics and reinforcement learning, this work provides a template for human-centered AI systems that improve through interaction while respecting individual differences.

Section 2 reviews related work on conversational surveys, personalization, and reinforcement learning for dialogue. Section 3 details the AURA architecture, LSDE scoring, state and action design, reward definition, and learning algorithm. Section 5 presents evaluation procedures and results. Section 6 concludes with implications and future directions.

2. Related Work

2.1. Conversational Surveys and Chatbot-Based Data Collection

AI-powered chatbots are increasingly used in higher education for information gathering, feedback collection, and student engagement. Institutions deploy them to collect course evaluations, campus-climate assessments, and service-quality feedback, often achieving higher response rates and more detailed input than traditional web forms (Abbas et al., 2021; Belhaj et al., 2021). For example, pilots at European universities show that conversational survey interfaces improve participation and perceived personalization (Martinez-Requejo et al., 2024). These deployments suggest that, when designed around user needs, chatbots can transform routine data collection into an engaging, interactive experience.

Within survey research, chat-based interfaces have emerged as an alternative to static questionnaires. Conversational surveys allow participants to respond in natural language and receive clarifying or follow-up questions in

real time (Xiao et al., 2020; Tanwar et al., 2024). Studies consistently report higher engagement and richer qualitative responses compared with conventional forms (Abbas et al., 2021; Kim et al., 2019). For instance, Xiao et al. (2020) found that participants produced more nuanced and emotionally expressive feedback when interacting with a chatbot, while Tanwar et al. (2024) demonstrated that large language models (LLMs) can generate context-aware follow-up questions for class feedback. At the same time, not all users prefer automated dialogue. Zarouali et al. (2024) and Njeguš et al. (2021) observed that some respondents still favored conventional surveys because of privacy or trust concerns, highlighting sociocultural factors that influence the acceptance of AI-mediated data collection.

The effectiveness of conversational surveys ultimately depends on the quality of the responses they elicit. Survey methodologists emphasize several indicators of response quality—length and elaboration (Dillman et al., 2014), emotional expression and engagement (Porter and Whitcomb, 2005), and the specificity of content (Brysbaert et al., 2014; Newman et al., 2003). Traditional questionnaires often fall short on these dimensions because they cannot adapt to respondent characteristics or probe for elaboration when needed. Chatbot interfaces offer the potential to improve response quality, but realizing this potential requires not only natural conversation but also strategic question selection that sustains engagement throughout the interaction.

Early chatbot survey systems were typically rule-based, relying on predefined dialogue trees and keyword matching (Thorat and Jadhav, 2020; Maeng and Lee, 2021; Chan et al., 2022). Such rigid architectures offered little flexibility: unexpected inputs often led to conversational dead ends or irrelevant prompts. Recent advances in natural language processing and LLMs have enabled greater contextual understanding, allowing chatbots to interpret ambiguous language and maintain coherent multi-exchanges (Xiao et al., 2020; Tanwar et al., 2024). This shift marks a transition from deterministic scripts to generative, data-driven dialogue. Nevertheless, even current LLM-based survey systems remain largely reactive. They generate follow-ups based on local context but do not learn from user behavior within a session or across interactions. Furthermore, most dialogue research focuses on task-oriented conversations where success is defined by goal completion (e.g., booking reservations or retrieving information) (Su et al., 2021). In contrast, survey quality depends on sustained engagement and willingness to elaborate, which cannot be reduced to a binary success metric.

This limitation is particularly evident in campus-climate surveys, where respondents vary widely in engagement patterns, disclosure comfort, and sensitivity to topics (Vogel et al., 2008; PASSHE, 2022). A first-year student discussing academic support may respond well to specific follow-up questions, whereas a faculty member describing discrimination may require validation and emotional acknowledgment before elaborating. Static prompts cannot accommodate such heterogeneity. While LLMs enable more natural language generation, they still lack a mechanism for *learning* which questioning strategies maintain engagement for each individual within a conversation. Addressing this limitation requires integrating conversational design with adaptive learning algorithms capable of optimizing question selection dynamically based on real-time feedback.

Our prior work, TigerGPT (Tang et al., 2025), demonstrated that generative chatbots incorporating empathetic cues, bolded questions, and user-driven topic selection can enhance satisfaction and produce richer feedback than static surveys in campus-climate assessment. However, TigerGPT—and most comparable systems—operated with fixed conversational heuristics: questions followed predetermined templates based on user role (student, faculty, or staff) but did not adapt to engagement patterns observed during the dialogue. These observations motivate the development of AURA, a reinforcement learning framework that achieves *within-session personalization* through continuous quality feedback and adaptive action selection.

2.2. Reinforcement Learning for Dialogue Management

Reinforcement learning (RL) provides a principled framework for optimizing sequential decisions in interactive systems where explicit supervision is limited. In dialogue management, RL formulates conversation as a Markov Decision Process (MDP) in which an agent selects actions—responses, question types, or dialogue strategies—based on conversational states to maximize cumulative reward such as task success or user satisfaction (Sutton et al., 1998). This paradigm enables systems to improve their policies through experience rather than relying solely on predefined rules.

Early research applied RL to task-oriented dialogues such as restaurant reservations or technical support. Li et al. (2016) pioneered deep RL for open-domain conversation, showing that policy-gradient optimization could enhance coherence and diversity beyond supervised models. Su et al. (2021) extended this framework through dynamic policy networks that adapt questioning strategies in real time for slot-filling tasks. Subsequent studies intro-

duced contextual bandits to improve personalization and sample efficiency (Sun et al., 2023), while recent work integrates RL with large language models via offline goal-conditioned learning to refine model policies without costly online trial-and-error (Hong et al., 2025). Together, these advances illustrate RL's potential for adaptive dialogue.

Within-session learning. Most RL dialogue systems employ multi-session learning: a global policy improves as the agent interacts with many users over time (Li et al., 2016; Su et al., 2021). Survey applications violate this assumption—each respondent interacts only once, exploration is costly, and privacy constraints limit cross-user adaptation (Kocaballi et al., 2019). *Within-session learning* instead focuses on rapid adaptation inside a single conversation. The agent begins from prior knowledge derived from prior data and adjusts its strategy in real time based on user engagement, prioritizing individual responsiveness over long-term population learning. Algorithms that converge within 10–15 exchanges are therefore required, motivating lightweight tabular or expected-value methods rather than deep networks.

Gap in existing work. Despite extensive work on RL for dialogue, prior systems optimize for task completion, coherence, or population-level personalization—not for maximizing individual response quality in brief, one-time interactions. Survey dialogues present unique challenges: continuous engagement rewards, limited interaction length, and substantial user heterogeneity. Addressing these requires (1) interpretable reward structures that quantify engagement improvements, (2) fast-adapting policies capable of learning within a single session, and (3) data-efficient value estimation to operate with limited prior samples. The AURA framework meets these needs through a reinforcement learning design that couples expected-value estimation with real-time quality feedback; full implementation details appear in Section 3.

2.3. Quality Assessment in Open-Ended Surveys

Prior research in survey methodology emphasizes that open-ended response quality is multidimensional rather than captured by a single indicator. Empirical and psycholinguistic studies identify four recurring factors: response length, self-disclosure, emotional expression, and specificity. Together, these features shape engagement and informativeness in qualitative data (Dillman et al., 2014; Holland and Christian, 2009; Pennebaker et al., 2003; Tausczik and Pennebaker, 2010; Brysbaert et al., 2014). These dimensions have been used in both human-coded (Krosnick, 1991; Holbrook et al., 2003) and automated text-analytic assessments. Recent conversational-survey

systems similarly exploit lexical and emotional cues to estimate engagement in real time (Xiao et al., 2020; Tanwar et al., 2024), motivating integration of such dynamic linguistic feedback within adaptive questioning frameworks. This body of work forms the theoretical foundation for AURA's linguistic quality metric introduced in Section 3.3, which operationalizes these dimensions for real-time evaluation within conversational surveys.

2.4. Individual Differences and the Need for Adaptive Questioning

Survey respondents vary widely in how they engage with different questioning strategies. Some provide rich detail when asked for specific examples, while others disengage or offer minimal replies; some respond well to emotional reflection prompts, whereas others prefer factual topic changes. This heterogeneity poses a fundamental challenge for conversational survey systems: approaches that succeed for one user can fail for another, and optimal strategies may shift even within a single conversation as topics change or respondent energy fluctuates. Adaptive systems must therefore learn individual preferences in real time rather than applying uniform strategies across all users.

Heterogeneity in survey response patterns. Research in survey methodology consistently demonstrates diverse engagement styles in open-ended responses. Krosnick (1991) identified systematic differences in response elaboration—some individuals naturally produce detailed narratives, while others provide brief replies even when equally knowledgeable. Holbrook et al. (2003) found that question-format effectiveness varies by respondent characteristics, with certain demographic groups responding better to specific than to general prompts. Tourangeau et al. (2000) showed that cognitive load, which fluctuates both across individuals and within conversations, affects how people process questions and formulate answers. Together, these findings suggest that optimal questioning complexity must adapt to each respondent's cognitive resources at any given moment.

Cognitive and affective factors. Several psychological mechanisms underlie these variations. Cognitive load theory (Lenzner et al., 2010) posits that question complexity interacts with working-memory capacity: respondents with lower available resources benefit from simpler, structured prompts, whereas those with higher capacity may disengage from overly basic questions. Affective factors further shape engagement. Melville and Hincks (2016) found that trust and willingness to self-disclose differ substantially across individuals and evolve throughout sensitive interviews. Some participants

9

require rapport building before sharing personal experiences, while others engage immediately. Wheeless and Grotz (1977) demonstrated that disclosure patterns are both person- and context-dependent, implying that effective interviewers must continuously calibrate their approach based on real-time feedback cues.

Communication styles and individual preferences. Individual differences also appear in preferred communication styles. Gudykunst and Ting-Toomey (1988) showed that people vary in receptiveness to direct versus indirect questioning, with some favoring explicit requests for information and others responding better to open-ended exploration. Preferences for elaborative versus succinct exchanges likewise differ: some respondents welcome follow-up probes that expand on previous answers, while others perceive repeated questioning as intrusive or redundant (Kim et al., 2019). Characteristics such as communication style, cultural background, educational experience, and topic familiarity all influence engagement patterns (Pennebaker and King, 1999; Gudykunst and Ting-Toomey, 1988), yet the specific combination varies by individual and context. Such variability indicates that effective conversational systems cannot rely on demographic stereotypes or population-level averages—they must infer user preferences dynamically during the interaction.

Evidence from conversational systems. Similar patterns appear in chatbot-based surveys. Zarouali et al. (2024) reported that response quality varied more by individual than by question type, underscoring the limitations of uniform dialogue strategies. In healthcare contexts, Kocaballi et al. (2019) showed that personalized conversational agents achieve higher satisfaction and greater information disclosure than generic systems, though most rely on explicit preference elicitation rather than real-time behavioral adaptation.

Limitations of one-size-fits-all approaches. Our pilot analysis provides direct evidence of these individual differences. The TigerGPT system (Tang et al., 2025) employed consistent conversational strategies across 96 campus-climate conversations, following empathetic design principles and user-driven topic selection. Yet among 467 responses collected, 64.2% exhibited low engagement (composite quality scores below 0.4). Response quality varied substantially across users: scores ranged from 0.0 to 1.0, with some respondents providing rich detail (above 0.6, 16.7 %) and others giving minimal one-word answers (below 0.2, 47.1 %). This heterogeneity demonstrates that even well-designed conversational systems with empathetic cues cannot succeed using fixed strategies. Effective adaptation requires learning individ-

ual preferences in real time from engagement patterns, motivating AURA's adaptive learning framework for discovering optimal questioning strategies within each single conversation.

2.5. Summary and Research Positioning

The related work reviewed above reveals a convergence of capabilities and constraints that motivate the design of AURA. Conversational survey systems achieve superior engagement compared with static questionnaires (Section 2.1), yet existing implementations rely on fixed questioning strategies that cannot adapt to individual users. Reinforcement learning offers a principled framework for adaptive dialogue (Section 2.2), but prior systems optimize for task completion or population-level learning across many users rather than maximizing individual response quality within single, brief interactions. Survey researchers have identified key response-quality dimensions—length, self-disclosure, emotion, and specificity (Section 2.3)—but these measures are typically assessed post-hoc through human coding instead of computed in real time for adaptive decision-making. Finally, extensive evidence documents individual differences in how respondents engage with questioning strategies (Section 2.4), yet no existing system learns optimal approaches for each user during their own conversation.

In response, AURA addresses these gaps through four integrated contributions. First, it operationalizes response quality as a composite real-time metric that combines length, self-disclosure, emotional intensity, and specificity, enabling automated assessment during conversation rather than relying on post-hoc human evaluation. Next, it employs reinforcement learning to optimize question selection based on observed quality improvements, treating engagement gains as reward signals that guide adaptive strategy choice. In addition, it implements *within-session learning*: the system begins each conversation with priors derived from 96 prior interactions but updates its estimates dynamically as each user responds, discovering personalized strategies within 10–15 exchanges. Finally, it integrates these components into a functioning conversational chatbot that combines the natural-language generation capabilities of large language models with the adaptive decision-making of reinforcement learning, enabling campus-climate surveys that adjust questioning strategies to each respondent's engagement patterns. Together, these elements establish AURA as the first framework to unify conversational empathy, quantitative quality metrics, and reinforcement learning for real-time adaptive surveying.

11

The remainder of this paper details AURA's implementation and evaluation. Section 3 describes the system architecture, including quality-scoring procedures, state representation, action selection, and reinforcement-learning algorithms. Section 5 presents empirical validation showing that AURA's adaptive approach yields significant quality improvements over non-adaptive baselines. Finally, Section 6 discusses the broader implications for conversational AI in sensitive organizational-assessment contexts and outlines directions for future research.

3. Methods

3.1. Overview of the AURA Framework

AURA (Adaptive Understanding through Reinforcement learning for Assessment) integrates conversational AI with reinforcement learning to enable real-time personalization in campus-climate surveys. Unlike traditional chatbots that follow fixed question sequences, AURA continuously monitors user engagement and adapts its questioning strategy within each conversation based on observed response-quality patterns. This adaptive loop transforms passive data collection into an interactive optimization process, allowing AURA to infer which questioning strategies maximize engagement within a single conversation and ultimately elicit richer, more detailed feedback than static survey instruments.

Figure 1 illustrates the complete architecture. The system operates as a continuous feedback loop: user responses are assessed for quality, generating reward signals that update action-selection policies in real time. This within-session learning enables the system to discover which questioning strategies work best for each individual participant.

AURA employs a two-level learning strategy that distinguishes it from conventional multi-session reinforcement learning systems (Figure 2). At the *offline level*, the system extracts patterns from 96 prior campus-climate conversations to initialize an expected-value (EV) table estimating the quality gain associated with each question type in each engagement state. This offline process establishes population-level priors derived from authentic survey interactions. At the *online level*, during each individual conversation, AURA updates its EV estimates in real time based on the current user's responses, discovering personalized strategies within 10–15 exchanges. Critically, the system resets to priors at the start of each new conversation—no informa-

12
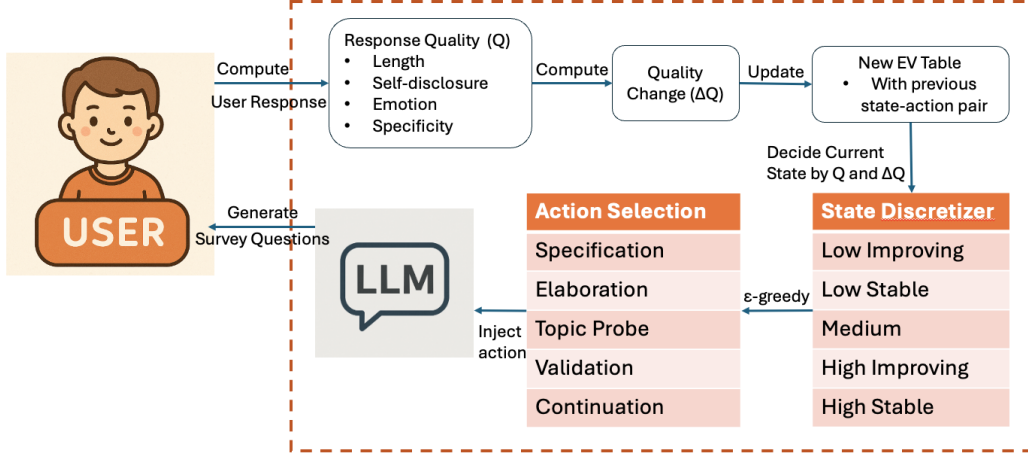
Figure 1: AURA system architecture showing the reinforcement learning cycle within a single conversation exchange. Each user response is scored along four quality dimensions (LSDE), mapped to an engagement state, and used to select the next question type via an ϵ-greedy policy. Observed quality changes update the system's expected-value (EV) estimates, enabling rapid within-session adaptation.

tion from previous users is retained—ensuring that adaptive learning remains session-specific and that participant privacy is fully preserved.

This architecture combines the generalizability of population-level patterns with the responsiveness of individual adaptation. By beginning each conversation with empirically grounded priors but adjusting strategy dynamically as user engagement evolves, AURA discovers personalized questioning approaches that static systems cannot achieve. The remainder of this section details the data sources (Section 3.2), quality-assessment methodology (Section 3.3), state and action representations (Sections 3.4–3.5), and reinforcement learning framework (Section 3.6).

3.2. Data Collection and Preprocessing

To initialize AURA's reinforcement-learning policy and establish empirical foundations for engagement measurement, we used campus-climate survey conversations collected from a pilot deployment of TigerGPT (Tang et al., 2025). TigerGPT is a semi-adaptive survey chatbot that conducts natural, open-ended dialogues but does not employ reinforcement learning. The re-
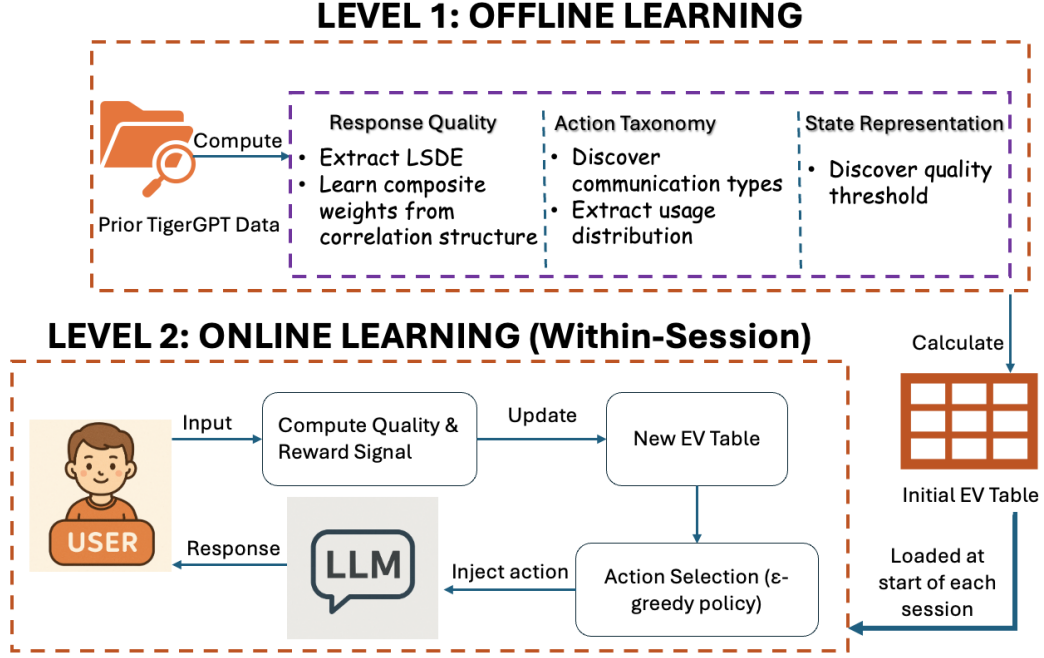
Figure 2: Two-level learning framework. Offline learning (top) extracts patterns from prior conversations to initialize the policy. Online learning (bottom) adapts to individual users within each session through real-time quality feedback, then resets to priors for the next user.

sulting dataset provides authentic examples of engagement in conversational surveys and enables offline estimation of action effectiveness across different engagement states.

3.2.1. Prior Conversation Dataset

The dataset comprises 96 campus-climate survey conversations conducted between October 2024 and February 2025 at the University of Missouri. The chatbot was publicly accessible through a web-based survey link, and all responses were collected anonymously as part of routine feedback activities. Each conversation included an initial role selection (student, faculty, or staff) followed by optional demographic prompts to personalize dialogue context. Participants then selected one of five available survey topics—academic issues, financial concerns, work–life balance, campus inclusivity, or a random option—and engaged in multi-exchange dialogues where TigerGPT posed

14

open-ended questions and users provided free-text responses.

3.2.2. Data Cleaning and Quality Control

From raw conversation logs totaling 683 exchanges, we applied standard preprocessing procedures: (1) removed entries where either the chatbot dialogue or user response was missing, empty, or contained placeholder strings ("nan", "N/A"); (2) excluded duplicate responses caused by technical errors; and (3) retained only substantive user inputs containing at least one word. This yielded a final dataset of 467 valid user responses across 96 conversations. The 31.6% exclusion rate reflects typical patterns in conversational data collection, where technical issues, participant dropout, or minimal engagement produce incomplete records.

3.2.3. Dataset Characteristics

Table 1: Summary Statistics of the Prior Conversation Dataset.

Category	Value
Conversation-level	
Total conversations	96
Mean exchanges per conversation	4.9 ± 5.5
Median exchanges per conversation	2.5
Range of exchanges	1–18
Single-exchange conversations	28 (29.2%)
Multi-exchange conversations (2+ exchanges)	68 (70.8%)
Response-level	
Total valid responses	467
Consecutive exchange pairs (for sequential analysis)	371
Mean response length (words)	18.3 ± 22.2
Median response length (words)	10.0

Table 1 summarizes the characteristics of the prior conversation dataset collected from TigerGPT, which serves as the empirical foundation for AURA's reinforcement learning initialization.

The dataset exhibits substantial heterogeneity in conversation length. Single-exchange interactions (29.2%) occurred when users provided initial feedback but chose not to continue; these responses contribute to cross-sectional quality measurement but cannot inform dynamics of engagement

change. Multi-exchange conversations (70.8%, $N = 68$) enable analysis of how response quality evolves across exchanges, producing 371 consecutive exchange pairs used for sequential analysis.

In this context, an *exchange* represents one full exchange between the chatbot and a user. An *exchange pair* links two consecutive user responses via the intervening chatbot question, enabling analysis of how question types affect response quality changes. Since the first response in each conversation has no preceding response for comparison, the 467 total responses yield 371 exchange pairs for sequential analysis ($467 - 96 = 371$).

The longest conversation reached 18 exchanges, though most remained brief (median $= 2.5$), reflecting typical engagement patterns in voluntary online surveys where respondents balance disclosure with time constraints. This distribution shaped AURA's design requirements: conversations are short enough (mean $= 4.9$ exchanges) that adaptive systems must learn rapidly, yet long enough (70.8% with multiple exchanges) to allow repeated state visits and within-session policy refinement.

The remainder of this section details how AURA transforms these prior conversations into an adaptive learning system: Section 3.3 presents the LSDE quality-scoring methodology applied to all responses, Section 3.4 defines the action space of question types, Section 3.5 describes the state representation that discretizes engagement patterns, and Section 3.6 explains the reinforcement learning algorithm that enables within-session personalization.

3.2.4. Data Usage in AURA Development

The prior conversation data served three primary functions in AURA's development.

Empirical grounding for quality metrics. The 467 user responses provided a reference distribution for establishing normalization parameters and threshold values in the LSDE scoring system (Section 3.3). For example, the length normalization cap was set at 29 words (the 75th percentile of the observed distribution), ensuring that normalized scores reflect empirical variability rather than arbitrary theoretical bounds.

Initial policy estimation. The 371 consecutive exchange pairs enabled the calculation of expected-value (EV) estimates for state–action combinations, capturing population-level patterns in how different question types affect engagement across distinct user states (Section 3.6). These estimates initialize AURA's policy before individual conversations begin, providing evidence-based starting points derived from authentic survey interactions

16

rather than random or uniform priors.

Action taxonomy validation. The diversity of question types in the corpus (Section 3.4) confirmed that the five-action categorization—*specification*, *elaboration*, *topic probe*, *validation*, and *continuation*—captured meaningful functional distinctions in conversational question design. Question classification used an LLM-based content analysis (GPT-4o) followed by human validation ($\kappa = 0.83$), establishing both the taxonomic structure and the empirical distribution of action usage in naturalistic campus-climate dialogues.

Importantly, these prior data inform AURA's initialization but do not constrain its behavior during live conversations. Each new user begins with population-level priors, but the system adapts its policy in real time based on individual response patterns, discovering personalized strategies that may diverge substantially from prior averages. The prior corpus thus serves as an empirically grounded prior rather than a fixed reference determining all subsequent decisions.

3.3. Response Quality Assessment (LSDE)

Effective conversational surveys require real-time measurement of response quality to guide adaptive question selection. We developed a four-dimensional quality metric (Length, Self-disclosure, Emotion, and Specificity; LSDE) that captures distinct aspects of engagement and enables automated assessment during live conversations. This section presents the theoretical foundations for each dimension, describes implementation procedures, and validates the composite scoring approach through empirical analysis of prior campus-climate conversations.

3.3.1. Foundations and Operationalization

Survey methodologists emphasize that response quality in open-ended questions is inherently multidimensional: no single indicator fully captures informativeness, engagement, or authenticity (Dillman et al., 2014; Tourangeau et al., 2000). Traditional assessment relies on post-hoc human coding, which cannot support real-time adaptive systems. AURA operationalizes four quality dimensions through automated linguistic analysis, enabling exchange-by-exchange measurement during conversation.

Length and Elaboration. Response length provides a first-order indicator of respondent investment and information content (Dillman et al., 2014; Holland and Christian, 2009). In traditional web surveys, detailed answers (≈ 50

17

words) are rated as substantive, but chatbot conversation exchanges are naturally shorter because users elaborate across multiple exchanges. In our dataset ($N = 467$), the median response length was 10 words and the 75th percentile (Q75) was 29 words, providing an empirically grounded scale for this dialogue context. Length is therefore normalized as

$$L_{\text{norm}} = \min\left(\frac{\text{word\_count}}{29}, 1.0\right) \tag{1}$$

so that the top quartile of responses receives full credit for elaboration while maintaining discrimination among briefer exchanges.

Self-Disclosure and Personal Engagement. First-person pronouns indicate self-focus and willingness to share personal experiences (Pennebaker et al., 2003; Tausczik and Pennebaker, 2010). In survey contexts, self-disclosure enriches qualitative feedback because responses grounded in personal experience (e.g., "I felt excluded when...") reveal deeper insight than abstract statements (e.g., "Some students experience exclusion"). Because our responses are brief (median $=$ 10 words), we use absolute rather than ratio-based pronoun counts to avoid mechanical dependence on response length: as text length increases, ratio scores decline even when disclosure remains constant, producing the strong Length–Disclosure correlation ($r = 0.856$; Table 2). Absolute counts, capped at the 75th-percentile value of three pronouns, measure disclosure directly while retaining discriminative range. Self-disclosure is therefore normalized as

$$D_{\text{norm}} = \min\left(\frac{\text{pronoun\_count}}{3}, 1.0\right) \tag{2}$$

where 3 represents the Q75 of our pronoun distribution. Pronouns include *I, me, my, mine, myself, we, us, our, ours, ourselves*, following LIWC's first-person category.

Emotional Expression and Intensity. Emotional language reflects psychological investment and topic salience. According to Russell's (1980) circumplex model of affect, emotion varies along valence (positive–negative) and arousal (intensity). In surveys, highly affective language—positive or negative—indicates engagement (Porter and Whitcomb, 2005). To measure this automatically, we apply VADER sentiment analysis (Hutto and Gilbert,

2014) and use the magnitude of its compound score as a proxy for emotional intensity, independent of polarity. Although VADER estimates valence rather than arousal directly, research on affective norms shows that extreme valence correlates moderately with arousal ($r \approx 0.40$; Warriner et al. 2013), making the absolute compound value a practical indicator of engagement strength in brief conversational responses. Emotional intensity is therefore normalized as:

$$E_{\text{norm}} = |\text{VADER}_{\text{compound}}| \tag{3}$$

This transformation yields normalized emotion scores on $[0, 1]$, where 0 indicates affectively neutral language and 1 indicates maximal emotional intensity.

Specificity and Concreteness. Concrete, detailed responses convey richer and more actionable information than vague generalities. Episodic memory research shows that specific recollection involves three components—entities (who/what), temporal references (when), and spatial references (where) (Levine et al., 2002). Authentic accounts contain more such contextual details than fabricated ones ($d = 0.72$; Newman et al. 2003). Unlike word-level concreteness metrics (Brysbaert et al., 2014), AURA detects the presence of these episodic components as binary indicators of specificity.

Each response is classified by GPT-4o for presence (1) or absence (0) of **entities**, **temporal**, and **spatial** details, following Levine et al.'s framework. Manual inspection of 50 samples confirmed consistent classification accuracy, after which the protocol was applied to the full corpus:

$$S_{\text{total}} = \text{entities} + \text{temporal} + \text{spatial} \in \{0, 1, 2, 3\}, \tag{4}$$

$$S_{\text{norm}} = \frac{S_{\text{total}}}{3}. \tag{5}$$

Specificity is extremely sparse in our data (88.7% of responses have $S_{\text{total}} = 0$, Q75 = 0). Although each component is detected binarily, their sum captures meaningful gradations: responses may contain one component (10.1%), two components (1.1%), or all three (0.2%). This count-based approach preserves distinctions among the 11.3% of responses containing specific detail while remaining robust to the severe sparsity.

3.3.2. Composite Scoring and Weighting Rationale

With each dimension normalized to $[0, 1]$, we combine them into a composite quality score. Equal weighting (0.25 per dimension) would overrep-

resent correlated features: Table 2 shows strong interdependencies among Length, Disclosure, and Emotion ($r = 0.52$–0.86), whereas Specificity is relatively independent (mean $|r| = 0.27$).

Table 2: Correlation Matrix of Normalized LSDE Dimensions ($N = 467$).

	Length	Disclosure	Emotion	Specificity
Length	1.000	0.856	0.664	0.290
Disclosure	0.856	1.000	0.523	0.248
Emotion	0.664	0.523	1.000	0.259
Specificity	0.290	0.248	0.259	1.000

Unequal weights were therefore assigned to balance three goals: (1) limit redundancy among highly correlated dimensions, (2) emphasize dimensions with broad empirical variation, and (3) retain all four theoretical constructs. The final composite metric is

$$Q_{\text{composite}} = 0.20 L_{\text{norm}} + 0.20 D_{\text{norm}} + 0.35 E_{\text{norm}} + 0.25 S_{\text{norm}} \qquad (6)$$

Weighting rationale.

- **Length and Disclosure (0.20 each; 0.40 total):** Highly correlated ($r = 0.86$); combined weight limits redundancy while reflecting their joint role in elaboration and engagement.

- **Emotion (0.35):** Moderately correlated ($r = 0.52$–0.66) but theoretically central to engagement intensity and empirically well distributed.

- **Specificity (0.25):** Statistically most independent but sparse (88.7 % zeros). A moderate weight preserves sensitivity for common responses while rewarding information-rich details.

3.3.3. Empirical Validation

To evaluate the composite metric, we computed quality scores for all 467 prior responses. Scores spanned the full $[0, 1]$ range (mean $= 0.328$, SD $= 0.276$), indicating broad variability and confirming that the metric captures meaningful differences in engagement levels (Table 3).

Nearly half of the responses (47.1%) fell into the Very Low category, reflecting the engagement challenges typical of campus-climate surveys in

Table 3: Distribution of Composite Quality Scores ($N = 467$).

Category	Score Range	n (%)
Very Low	0.0–0.2	220 (47.1%)
Low	0.2–0.4	80 (17.1%)
Medium	0.4–0.6	90 (19.3%)
High	0.6–0.8	70 (15.0%)
Very High	0.8–1.0	7 (1.5%)

which many participants provide only minimal input (e.g., one-word replies). This distribution demonstrates that the LSDE metric can distinguish low from high engagement with fine granularity, supporting its use as a reward signal for reinforcement learning models that aim to improve response quality through adaptive question selection.

3.4. Action Space Extraction

Effective reinforcement learning for conversational surveys requires a well-defined *action space* representing the questioning strategies available to the system. Rather than imposing predetermined categories, we extracted action types empirically from prior TigerGPT conversations and mapped them to established communication theory frameworks. A GPT-4o–based content analysis labeled each survey question into one of five communicative functions—*specification, elaboration, topic probe, validation,* and *continuation.* The model prompt instructed classification by primary and optional secondary intents, with structured JSON outputs including confidence and brief reasoning. Human review of a random subset confirmed consistent labeling ($\kappa = 0.83$). This process yielded a compact yet comprehensive taxonomy suitable for reinforcement-learning–based policy optimization.

3.4.1. Taxonomy Development: A Theory-Informed, Data-Driven Approach

Action taxonomy development followed a hybrid methodology balancing empirical observation with theoretical grounding. Analysis of 467 chatbot questions across 96 TigerGPT conversations (Section 3.2) revealed recurring functional patterns: some prompts requested concrete examples, others sought elaboration, introduced new topics, or offered acknowledgment. These patterns were consolidated into five communicative categories and validated

against established communication frameworks to ensure conceptual coherence (Li et al., 2016; Su et al., 2021).

This hybrid strategy avoids the limitations of purely data-driven extraction—which can yield uninterpretable clusters—and purely theory-driven design—which may impose artificial categories unrelated to the data. AURA's action space is thus both **empirically grounded** (all categories appear with sufficient frequency in natural dialogues) and **theoretically justified** (each aligns with established constructs in communication and dialogue management).

The five-category structure balances interpretability with statistical power. With 96 conversations producing 371 consecutive exchange pairs, compound multi-intent schemes would create extreme data sparsity. For instance, allowing two-intent combinations would expand the action space from 5 to 15 categories, and three-intent combinations to 25—yielding only about 15 examples per category on average, with many categories represented by fewer than five instances. AURA therefore uses primary-intent classifications for reinforcement learning while retaining secondary intents for post-hoc analysis, prioritizing statistical robustness over taxonomic exhaustiveness.

3.4.2. Action Classification Methodology

Each chatbot dialogue in the prior corpus was classified by GPT-4o using a structured prompt defining five communicative categories and specifying primary versus secondary intents. The model identified the main communicative function of each question (primary intent) and any auxiliary functions (secondary intents). Approximately 65 % of questions exhibited multiple intents (e.g., "Thank you for sharing. Could you tell me more about that specific example?" combines *validation* and *specification*), illustrating the prevalence of compound strategies in natural dialogue.

Primary intent was defined as the question's dominant illocutionary force (Searle, 1976)—the central action the question seeks to elicit. Model outputs included structured JSON records with confidence scores and brief reasoning to support quality control and flag ambiguous cases.

To validate classification accuracy, a random sample of 20 questions was human-coded, showing strong agreement between GPT-4o labels and theoretical definitions. This human-in-the-loop check confirmed that the automated classification captured meaningful functional distinctions rather than superficial lexical cues.

Table 4: Action Type Taxonomy, Theoretical Grounding, and Prior Examples

Action Type	Speech Act Function	CAT Strategy	Example
specification	Directive	Convergence (narrowing focus)	"Could you share how satisfied you are with your academic program and why?"
elaboration	Directive	Maintenance (sustaining topic)	"Could you share more about the challenges you've faced in Greek life?"
topic probe	Commissive	Divergence (expanding scope)	"How comfortable do you feel on campus, and have you experienced moments of inclusion or exclusion?"
validation	Expressive	Maintenance (rapport building)	"Thank you for sharing that. Your perspective is valuable and helps us understand student experiences."
continuation	Directive	Maintenance (flow sustaining)	"Is there anything else you'd like to share about your experiences at MU?"

Note: CAT = Communication Accommodation Theory (Giles et al., 1991).
Speech act categories follow Searle (1976)'s taxonomy of illocutionary forces.

3.4.3. Action Type Definitions and Theoretical Grounding

Table 4 presents the five action types with their theoretical foundations and representative examples from the TigerGPT corpus. Each action corresponds to distinct constructs from Speech Act Theory (Searle, 1976) and Communication Accommodation Theory (CAT) (Giles et al., 1991), ensuring that AURA's taxonomy captures functionally meaningful communicative strategies rather than superficial lexical patterns.

Specification and **elaboration** both operate as directive speech acts but differ in focus. Specification requests concrete examples or particular cases, using CAT's convergence strategy to narrow discussion toward actionable details (e.g., transforming "Classes can be challenging" into "Professor Smith's EECS 280 lectures often run overtime"). Elaboration maintains the current topic while seeking richer description without requiring specific instantiation (e.g., "Tell me more about your research experience"). The distinction is functional: specification targets concreteness, elaboration depth.

Topic probe introduces new experiential dimensions, shifting attention to unexplored areas. These prompts function at the intersection of commissive and directive acts (Searle, 1976), committing the system to expanding the dialogue scope. They reflect CAT's divergence strategy (Giles et al., 1991), ensuring broad coverage of campus climate dimensions (academic, social, financial, inclusivity).

Validation acknowledges user contributions without requesting new information. As expressive acts, validation supports rapport and psychological safety through maintenance strategies (Giles et al., 1991). Although rare in prior data ($n = 4$, 0.9%), validation was retained because evidence shows it facilitates deeper disclosure on sensitive topics (Wheeless and Grotz, 1977; Melville and Hincks, 2016).

Continuation invites further input without specifying direction, granting users conversational autonomy. Also infrequent ($n = 2$, 0.4%), continuation serves as a flow-maintaining strategy that may prove effective for users preferring open-ended engagement (Kocaballi et al., 2019).

3.4.4. Action Distribution in Prior Data

Table 5 summarizes the distribution of action types in the prior TigerGPT corpus. The data show a pronounced imbalance: *specification* dominates (62.3%), while *validation* (0.9%) and *continuation* (0.4%) are rarely used. This pattern reflects conventional survey design practices that prioritize information extraction over rapport maintenance or open-ended exploration.

Table 5: Action Type Distribution in Prior Corpus ($N = 467$ conversation exchanges)

Action Type	Count	Percentage (%)	Characteristics
Specification	291	62.3	Requests concrete examples
Elaboration	110	23.6	Asks for expanded detail
Topic probe	60	12.8	Introduces new dimensions
Validation	4	0.9	Acknowledges contributions
Continuation	2	0.4	General follow-up
Total	**467**	**100**	

Note: Distribution reflects prior TigerGPT behavior. AURA's RL framework enables adaptive rebalancing for individual users.

From a reinforcement learning perspective, this imbalance presents both a challenge and an opportunity. The challenge is data sparsity: rare actions have too few examples for reliable expected-value (EV) estimation, creating cold-start conditions for several state–action pairs. Specifically, eight of twenty-five combinations lack any prior samples (Section 3.6), requiring AURA to explore them during live interactions. The opportunity lies in AURA's capacity to **correct prior biases**: the dominance of specification likely reflects template-driven defaults rather than optimal engagement strategies. By adaptively testing underused actions such as *validation* after emotional responses or *topic probe* during stagnation, AURA can discover more effective questioning patterns for individual users.

An ϵ-greedy exploration policy ($\epsilon = 0.30$; Section 3.6) ensures all actions receive exposure despite imbalance. Roughly 30% of actions per conversation are randomly selected, allowing even rare strategies to be evaluated. When these actions improve subsequent response quality, their expected values increase and they are chosen more often for that user; conversely, ineffective actions decline in frequency. Through this continual rebalancing, AURA learns personalized questioning strategies that static, rule-based systems cannot achieve.

3.4.5. Multi-Intent Analysis and Primary Action Selection

About 65 % of prior questions contained multiple intents, showing that conversational prompts often blend functions. For instance, "Thank you for sharing that. Could you tell me about a particular instance when this happened?" combines *validation* with *specification*, illustrating how rapport

building and information elicitation frequently co-occur.

For reinforcement learning, we label each question by its **primary intent**—the dominant communicative function. This choice resolves a key design trade-off: compound actions capture conversational nuance but generate extreme data sparsity. With five base types, even two-intent combinations expand the action space to 15 categories and three-intent combinations to 25+, leaving fewer than five samples per class across 96 conversations—insufficient for reliable expected-value estimation.

Using primary intent yields a tractable five-action space (25 state–action pairs) while retaining compound information for post-hoc analysis. Secondary intents were recorded but excluded from training, allowing later evaluation of whether particular combinations (e.g., *specification + validation*) consistently enhance engagement. This design favors **within-session learning feasibility** over exhaustive taxonomy: AURA must learn effective strategies within 10–15 exchanges per user. Future extensions could explore hierarchical policies that first select a primary action, then conditionally choose secondary intents as data volume grows.

3.5. State Representation Design

Having established AURA's five-action question taxonomy, we now define the state representation that captures user engagement patterns during conversation. Effective reinforcement learning requires discrete states that both reflect meaningful engagement differences and enable sufficient data for policy learning within short conversations. This section describes the state discretization approach, threshold selection rationale, and empirical validation of the five-state framework.

3.5.1. Design Requirements and Constraints

State representation in conversational reinforcement learning must balance expressiveness with learnability. AURA faces two critical constraints that shape state design. First, conversations are brief: our prior conversations exhibit a mean of 4.9 exchanges per conversation (median = 2.5), and 70.8% contain at least two exchanges. Adaptive systems must therefore learn rapidly, discovering effective strategies within the limited exchanges typical of voluntary survey interactions. Second, AURA employs session-specific learning with no cross-user memory retention: the system initializes each conversation with population-level priors but adapts only to the current user,

resetting completely before the next interaction. This design preserves privacy and prevents overfitting to specific individuals but requires state spaces that enable meaningful learning from limited within-session experience.

These constraints necessitate discrete rather than continuous state representations. In a typical conversation with continuous states (where each quality score is unique), the system would encounter each state exactly once, precluding any learning from repeated state visits. Discrete states, by contrast, enable state revisits: if a user's engagement remains in the "medium" category across multiple exchanges, the system observes which actions succeed or fail in that state, updating its policy accordingly. With 5 discrete states, AURA is designed for conversations of 10-15 exchanges, enabling users to revisit each state 2-3 times on average and providing sufficient feedback for within-session adaptation. Although prior TigerGPT conversations averaged only 4.9 exchanges, AURA's adaptive design aims to sustain longer, more productive dialogues.

3.5.2. Hybrid State Representation: Quality Level and Trajectory

We adopt a hybrid approach that combines absolute engagement level with recent trajectory, yielding five states: `low_improving`, `low_stable`, `medium`, `high_improving`, and `high_stable`. Each state is determined by two factors: the current response's composite quality score ($Q_{\text{composite}}$) and the change in quality relative to the previous exchange (ΔQ):

$$\Delta Q_t = Q_{\text{composite},t} - Q_{\text{composite},t-1} \tag{7}$$

State assignment follows a hierarchical decision rule. First, quality is categorized into three levels based on thresholds derived from the prior data distribution: low ($Q < 0.3$), medium ($0.3 \leq Q < 0.6$), and high ($Q \geq 0.6$). These thresholds were selected to align with meaningful quality distinctions: the low threshold (0.3) approximates the dataset mean (0.328), capturing below-average responses; the high threshold (0.6) exceeds the 75th percentile ($Q75 = 0.520$), identifying top-quartile engagement.

For low and high quality levels, trajectory information further refines the state. A quality change exceeding $|\Delta Q| > 0.05$ indicates improving engagement, while smaller changes ($|\Delta Q| \leq 0.05$) reflect stable patterns. The 0.05 threshold was calibrated to detect meaningful shifts without excessive sensitivity to measurement noise: it represents approximately one standard deviation of exchange-to-exchange variability in the prior corpus.

Responses in the medium quality range are not subdivided by trajectory, as pilot analysis revealed insufficient distinction between improving and stable medium-engagement patterns given the available data.

The complete state assignment function is:

$$s_t = \begin{cases} \texttt{low\_improving} & \text{if } Q_t < 0.3 \text{ and } \Delta Q_t > 0.05 \\ \texttt{low\_stable} & \text{if } Q_t < 0.3 \text{ and } \Delta Q_t \leq 0.05 \\ \texttt{medium} & \text{if } 0.3 \leq Q_t < 0.6 \\ \texttt{high\_improving} & \text{if } Q_t \geq 0.6 \text{ and } \Delta Q_t > 0.05 \\ \texttt{high\_stable} & \text{if } Q_t \geq 0.6 \text{ and } \Delta Q_t \leq 0.05 \end{cases} \tag{8}$$

For the first exchange of each conversation, where no previous quality score exists, ΔQ is defined as zero. This assigns the user to either a "stable" state (`low_stable` or `high_stable`) or to `medium`, depending solely on initial quality level.

3.5.3. Rationale for Quality-Trajectory Hybrid

The hybrid design addresses a fundamental question in adaptive surveying: should the system's next action depend only on how engaged the user currently is, or also on whether engagement is increasing or decreasing? Pure quality-level states (e.g., simply "low," "medium," "high") ignore dynamics, treating a user whose engagement is declining identically to one whose engagement is improving—even though these patterns may warrant different questioning strategies. Conversely, pure trajectory-based states (e.g., "improving," "stable," "declining") disregard absolute engagement levels, potentially recommending the same action for a slightly improving disengaged user and a slightly improving highly engaged user.

The hybrid approach captures both dimensions. For example, `low_improving` indicates rising quality from a low baseline, suggesting that recent question types may be resonating with the user. By contrast, `low_stable` reflects persistent low engagement, indicating that previous strategies have not improved response quality. Similarly, distinguishing `high_improving` from `high_stable` enables the system to differentiate between users building momentum and those maintaining consistent engagement—patterns that may respond differently to continued questioning. This two-factor representation provides the reinforcement learning algorithm (Section 3.6) with sufficient information to adapt its strategy based on both current engagement and recent trajectory.

28

3.5.4. Empirical Validation and Coverage

To validate the five-state framework, we categorized all 467 prior responses according to Equation 8. Table 6 presents the resulting distribution. The `low_stable` state dominates (46.7%), reflecting the engagement challenge inherent in voluntary campus-climate surveys: many participants provide minimal responses. The `medium` state accounts for 27.8% of responses, representing moderate engagement. High-engagement states (`high_improving` and `high_stable`) together constitute 16.7%, confirming that sustained detailed disclosure remains relatively rare in brief survey interactions.

Table 6: State Distribution in Prior Conversation Data (N = 467 responses).

State	Count	Percentage (%)	Mean Quality Score
low_stable	218	46.7	0.071
medium	130	27.8	0.454
high_improving	60	12.8	0.709
low_improving	41	8.8	0.184
high_stable	18	3.9	0.738
Total	467	100.0	0.328

Critically, all five states are well-represented in the prior data, ensuring sufficient examples for policy initialization (Section 3.6). Although `high_stable` constitutes only 3.9% of responses—the smallest category—this proportion still represents 18 observations, adequate for establishing baseline patterns while acknowledging that rare states will benefit most from within-session adaptation as the system encounters individual users.

The state categories exhibit meaningful quality separation. Mean composite scores range from 0.071 (`low_stable`) to 0.738 (`high_stable`), spanning nearly the full measurement range and confirming that the discretization captures substantive engagement differences rather than arbitrary divisions. The trajectory distinction within low and high ranges further refines these categories: `low_improving` (mean = 0.184) falls between `low_stable` (0.071) and `medium` (0.454), consistent with users transitioning from disengagement toward moderate participation. This empirical validation confirms that the five-state representation provides both sufficient coverage of engagement patterns and meaningful differentiation for reinforcement learning.

3.6. Reinforcement Learning Framework

Having established AURA's quality-scoring methodology (Section 3.3), action taxonomy (Section 3.4), and state representation (Section 3.5), we present the reinforcement learning algorithm enabling within-session adaptation. AURA employs a two-level architecture (Figure 2): offline learning initializes an expected-value (EV) table from 96 prior conversations; online learning adapts the policy in real time based on each user's responses. This section describes both phases and details question generation.

3.6.1. Offline Learning: Initial Expected-Value Calculation

Before engaging with users, AURA initializes its policy using expected-value estimates derived from 371 prior exchange pairs (Section 3.2). Unlike frequency-based approaches that select the most common actions, AURA's EV estimation quantifies each action's effectiveness by combining success probability and improvement magnitude.

Expected-Value Formula. For each state-action pair (s, a):

$$\text{EV}(s, a) = P(\text{improvement} \mid s, a) \times \mathbb{E}[\Delta Q \mid \text{improvement}, s, a] \qquad (9)$$

where:

- $P(\text{improvement} \mid s, a)$ is the proportion of instances where action a in state s resulted in quality improvement $(\Delta Q > 0)$

- $\mathbb{E}[\Delta Q \mid \text{improvement}, s, a]$ is the average quality gain when improvement occurred

- $\Delta Q = Q_{\text{composite},t+1} - Q_{\text{composite},t}$ is the change in composite quality between consecutive exchanges

This formulation captures both the probability and magnitude of quality improvements.

Calculation Example. For the state-action pair (*low_stable*, *topic probe*):

- Observations: 20 exchange pairs

- Improvements: 16 of 20 instances (80%)

- Average ΔQ when improved: +0.3815

Table 7: Initial Expected-Value Estimates from Prior Data (N = 371 exchange pairs)

State	Action	EV	n	Confidence
low_improving	specification	0.058	15	Moderate
	elaboration	0.047	9	Moderate
	topic_probe	0.032	3	Low[†]
	validation	0.000	0	None[‡]
	continuation	0.000	0	None[‡]
low_stable	specification	0.288	112	**Reliable**
	elaboration	0.170	27	**Reliable**
	topic_probe	0.305	20	**Reliable**
	validation	0.348	4	Low[†]
	continuation	0.476	1	Low[†]
medium	specification	0.071	66	**Reliable**
	elaboration	0.073	28	**Reliable**
	topic_probe	0.039	22	Moderate
	validation	0.000	0	None[‡]
	continuation	0.000	0	None[‡]
high_improving	specification	0.004	33	Moderate
	elaboration	0.020	14	Moderate
	topic_probe	0.000	4	Low[†]
	validation	0.000	0	None[‡]
	continuation	0.000	0	None[‡]
high_stable	specification	0.040	9	Moderate
	elaboration	0.083	1	Low[†]
	topic_probe	0.028	3	Low[†]
	validation	0.000	0	None[‡]
	continuation	0.000	0	None[‡]

Note: Confidence categories: **Reliable** ($n \geq 20$), Moderate ($5 \leq n < 20$), Low[†] ($1 \leq n < 5$), None[‡] ($n = 0$, cold-start pairs). EV values represent expected quality improvement; higher values indicate more effective actions for that state.

- $\mathrm{EV}(\mathrm{low\_stable}, \mathrm{topic\_probe}) = 0.80 \times 0.3815 = 0.305$

Table 7 presents the complete set of initial EV estimates with sample sizes and confidence indicators. Seventeen of 25 state-action pairs (68%) have prior data; the remaining 8 pairs (32%) are initialized to EV = 0.0, representing neutral prior expectation.

Coverage and Confidence. Among pairs with data, five have $n \geq 20$ (reliable estimates), six have $5 \leq n < 20$ (moderate confidence), and six have $n < 5$ (low confidence). The exploration rate ($\epsilon = 0.30$) described in Section 3.6.2 ensures that all actions, including those with limited or no prior data, receive trials during individual conversations.

3.6.2. Online Learning: Within-Session Adaptation

During each conversation, AURA updates its policy in real time based on observed quality changes, enabling rapid personalization. This within-session learning combines an ϵ-greedy policy balancing exploration and exploitation with an update rule that adjusts EV estimates after each exchange.

Reward Signal. At exchange t, the immediate reward is the change in composite quality:

$$r_t = \Delta Q_t = Q_{\mathrm{composite},t} - Q_{\mathrm{composite},t-1} \tag{10}$$

For $t = 1$, no previous quality exists, so r_1 is undefined and no update occurs.

ϵ-greedy Action Selection. At exchange t, given state s_t, AURA selects the next action:

$$a_t = \begin{cases} \text{random action from } \mathcal{A} & \text{with probability } \epsilon = 0.30 \\ \arg\max_{a \in \mathcal{A}} \mathrm{EV}(s_t, a) & \text{with probability } 1 - \epsilon = 0.70 \end{cases} \tag{11}$$

The system explores randomly 30% of the time, ensuring all actions receive trials despite prior imbalance (specification 62.3%; validation 0.9%, continuation 0.4%). Otherwise, it exploits current knowledge by selecting the highest-EV action.

EV Update Rule. After observing reward r_t, AURA updates the expected value for the state-action pair from the previous exchange:

$$\mathrm{EV}(s_{t-1}, a_{t-1}) \leftarrow \mathrm{EV}(s_{t-1}, a_{t-1}) + \alpha \left[r_t - \mathrm{EV}(s_{t-1}, a_{t-1}) \right] \tag{12}$$

where $\alpha = 0.3$ is the learning rate. This standard temporal-difference update (Sutton et al., 1998) occurs at exchange t after observing the outcome of

32

action a_{t-1} selected in state s_{t-1} at the previous exchange, increasing EV when the observed reward exceeds the prior estimate and decreasing it when the action underperforms.

These EV updates are session-specific: each conversation initializes with the prior EV table (Table 7), adapts based on the current user's responses, then discards the modified table at session end, ensuring privacy and individual-focused adaptation.

Exchange-by-Exchange Process. At each exchange t, AURA executes the following sequence:

1. Collect and score user response: Compute Q_t via LSDE (Section 3.3)
2. If $t > 1$:
    - Compute reward: $r_t = Q_t - Q_{t-1}$
    - Update previous action's value: Apply Equation 12 to (s_{t-1}, a_{t-1})
3. Discretize current state: Map $(Q_t, \Delta Q_t) \to s_t$ via Section 3.5
4. Select next action: Choose a_t using ϵ-greedy policy (Equation 11)
5. Generate question: Produce natural-language prompt (Section 3.6.3)

After 15 exchanges or user termination, the session ends and all within-session EV updates are discarded, ensuring the next user begins fresh with priors.

3.6.3. Question Generation

The reinforcement learning module selects abstract action types, but users interact with natural-language questions. AURA employs GPT-4o-mini to translate RL decisions into contextually appropriate prompts.

Action-to-Prompt Mapping. Each action type (Section 3.4) is associated with generation directives that instruct the LLM:

- Specification: Request specific examples with contextual details

- Elaboration: Request deeper explanation about reasoning or feelings

- Topic Probe: Explore related but different aspects of campus life

- Validation: Briefly acknowledge contributions (under 20 words)

- Continuation: Provide minimal prompt (5-10 words) to sustain flow

33

Context Integration. The prompt includes conversation history (previous 2–3 exchanges) and action-specific instructions. For specification and elaboration, the LLM deepens the current topic; for topic probe, it explores related campus climate dimensions (academic, social, diversity, resources, mental health) while maintaining conversational coherence.

Generation Parameters. The LLM uses temperature $= 0.7$ to generate natural variation in phrasing while maintaining contextual appropriateness. The system does not employ strict templates; each question is generated dynamically based on conversation history and selected action type.

Example. Given action $=$ *topic probe* after discussion about Greek life exclusivity, the system generates: *"That's a really important perspective. Beyond Greek life, how would you describe the broader sense of community and belonging you feel on campus—like in classes, clubs, or other social settings?"* The question transitions to related content (general social climate) through explicit connection to the previous topic.

By separating strategic decisions (which action) from tactical implementation (how to phrase), the system combines RL optimization with LLM fluency.

4. Evaluation Design

4.1. Research Questions & Hypotheses

This evaluation validates AURA's within-session learning capability and optimizes the exploration parameter (ϵ) for conversational survey applications. Analysis of the prior TigerGPT dataset (Section 3.2) revealed problematic questioning patterns: over-reliance on specification (62.3%, Table 5) and severe under-utilization of validation (0.9%). These patterns reflect common practices in conversational survey design where systems prioritize information extraction over rapport maintenance. We address two research questions:

RQ1: Does AURA's RL-based adaptation improve response quality compared to prior TigerGPT behavior?

We test whether within-session learning produces significant quality improvements ($p < 0.05$, Cohen's $d > 0.50$) over the non-adaptive prior baseline, measured by quality change from initial to final conversation exchanges.

RQ2: What exploration rate (ϵ) optimizes within-session learning in brief survey conversations?

We compare three ϵ-greedy configurations to identify the exploration-exploitation balance that maximizes quality gains within 15-exchange in-

teractions: conservative fixed ($\epsilon = 0.15$), moderate fixed ($\epsilon = 0.30$), and decaying ($\epsilon : 0.40 \to 0.05$).

4.2. Baseline System

The **prior distribution baseline** replicates actual TigerGPT behavior from the dataset described in Section 3.2. This semi-adaptive campus climate survey system, deployed October 2024–February 2025, collected the 96 conversations (371 consecutive exchange pairs) used to initialize AURA's expected-value table (Section 3.6). While TigerGPT employed contextual follow-ups and flexible topic selection, it lacked reinforcement learning and exhibited the action distribution patterns shown in Table 5: specification 62.3%, elaboration 23.6%, topic probe 12.8%, validation 0.9%, and continuation 0.4%.

The baseline implements these empirical action frequencies through weighted random sampling. At each exchange, actions are selected according to the prior probabilities, independent of current engagement state, conversation history, or observed quality patterns. The baseline exhibits **no state awareness**—all exchanges use identical probabilities regardless of user engagement level—and **no learning mechanism**—the distribution remains constant throughout and across conversations.

This baseline operationalizes two engagement problems AURA targets: (1) **specification over-reliance** at 62.3% risks respondent fatigue through repetitive requests for concrete examples, and (2) **validation neglect** at 0.9% undermines rapport-building despite its theoretical importance for sustaining engagement. Improvement over this ecologically valid baseline—representing real survey practice rather than an artificial control—demonstrates that RL adaptation addresses documented methodological challenges in conversational data collection.

4.3. Evaluation Metrics

Primary metric: Quality Improvement

$$\Delta Q = Q_{\text{final}} - Q_{\text{initial}} \tag{13}$$

where Q_{initial} and Q_{final} are composite LSDE scores at Exchange 1 and Exchange 15. This metric isolates adaptive contribution by controlling for baseline engagement, directly measuring AURA's core objective of improving response quality through within-session learning.

Secondary metrics: Final quality (Q_{final}), average quality ($\bar{Q}$), and exchange success rate (proportion of exchanges with $\Delta Q > 0$).

Action-level analysis: Distribution of five action types and action appropriateness (proportion of selections with positive expected value in the prior initialization table, Table 7).

Temporal analysis: Early (Exchange 1–5), mid (6–10), and late (11–15) quality trajectories to characterize adaptation dynamics across conversation phases.

4.4. Experimental Design

We employed a between-subjects factorial design with 4 system conditions (3 RL variants + prior baseline) × 4 simulated user profiles. Each condition comprised $n = 20$ conversations (5 repetitions per profile), yielding 80 total conversations. All RL configurations shared $\alpha = 0.30$ learning rate, 15-exchange conversations, and within-session learning only (complete EV table reset between users to preserve privacy and ensure individual-focused adaptation).

RL Exploration Configurations. Three epsilon variants tested alternative exploration-exploitation hypotheses:

- **Configuration 1 ($\epsilon = 0.15$):** Conservative fixed exploration with steady 15% random action sampling throughout all exchanges. The hypothesis was that low exploration prevents distraction from learned patterns while allowing occasional discovery of alternatives.

- **Configuration 2 ($\epsilon = 0.30$):** Moderate fixed exploration with consistent 30% random sampling. This higher exploration rate was hypothesized to prevent premature convergence to suboptimal policies, particularly important given sparse prior data (only 17 of 25 state-action pairs have examples, Section 3.6).

- **Configuration 3 ($\epsilon : 0.40 \to 0.05$):** Decaying exploration schedule with high initial exploration (40%) decreasing linearly to minimal exploration (5%) by Exchange 15. At each exchange t, $\epsilon_t = 0.40 - (0.35 \times t/15)$. The hypothesis was that early exploration discovers effective actions while late exploitation refines learned strategy.

Simulated User Profiles. To enable experimental control, reproducibility, and statistical power while maintaining computational feasibility, we

36

employed AI-simulated users rather than human participants. Simulated users were implemented using GPT-4o-mini (temperature $= 0.8$) with profile-specific prompts ensuring consistent persona characteristics. Four profiles were selected to represent diverse disciplinary backgrounds and academic levels commonly found in campus populations:

- **Biology Senior:** Science-oriented student with moderate engagement style, focused on academic experiences and research opportunities.

- **Psychology Junior:** Social science perspective with high emotional expressiveness and interpersonal focus.

- **Computer Science Sophomore:** Technical background with analytical communication style and emphasis on academic rigor.

- **English Senior:** Humanities orientation with elaborative responses and attention to campus culture and community.

This disciplinary diversity (STEM, social science, humanities) combined with variation in academic seniority (sophomore, junior, senior) ensured that evaluation captured a range of communication styles and engagement patterns typical of university survey populations. Profile characteristics remained constant across all system conditions, ensuring fair comparison and isolating the effect of questioning strategy from individual differences.

The use of simulated users enabled three critical experimental advantages: (1) precise control, with identical profiles tested across all conditions; (2) perfect reproducibility, with conversations deterministically reconstructable from logs; and (3) adequate statistical power (> 0.80 for $d \geq 0.65$ at $\alpha = 0.05$, given $n = 20$ per condition). While simulated users cannot capture the full complexity of human behavior, they provide a controlled environment for validating within-session learning mechanisms before deployment with actual survey participants.

Statistical Analysis. We employed independent samples t-tests comparing each RL configuration to the prior baseline. Three planned comparisons were conducted with uncorrected p-values reported ($\alpha = 0.05$, two-tailed). Cohen's d effect sizes were computed for all comparisons, with interpretative guidelines: small ($d = 0.20$–0.49), medium ($d = 0.50$–0.79), large ($d \geq 0.80$).

5. Results

5.1. Overall Performance Comparison

Table 8 presents primary results comparing three RL configurations to the prior baseline. Configuration 2 ($\epsilon = 0.30$) significantly outperformed the baseline on quality improvement ($t(38) = 2.088$, $p = 0.044$, $d = 0.660$), demonstrating that moderate fixed exploration enables effective within-session adaptation. Configuration 2 also achieved substantially higher final quality scores ($Q_{\text{final}} = 0.582$ vs. baseline 0.465, a difference of $+0.117$). Configuration 1 ($\epsilon = 0.15$) showed marginal improvement ($p = 0.078$, $d = 0.573$), while Configuration 3 (decaying epsilon) failed to reach significance ($p = 0.345$, $d = 0.302$).

The results provide clear answers to our research questions. For **RQ1**, RL-based adaptation yields significant quality improvements over non-adaptive baseline behavior when properly configured. The best-performing system (Configuration 2) achieved a quality improvement advantage of $+0.076$ relative to baseline (effect size $d = 0.660$, medium effect), and final quality exceeded baseline by $+0.117$. These effects demonstrate that within-session learning successfully addresses the engagement problems documented in the prior data.

For **RQ2**, moderate fixed exploration ($\epsilon = 0.30$) proved optimal for brief survey conversations. Configuration 1's conservative exploration ($\epsilon = 0.15$) approached but did not achieve significance, suggesting insufficient sampling leads to premature convergence (examined further in Section 5.2). Configuration 3's decaying schedule ($\epsilon : 0.40 \rightarrow 0.05$) produced the weakest performance, indicating that high early exploration introduces excessive variability that disrupts learning before the system can benefit from increased exploitation.

Notably, the prior baseline exhibited minimal quality change ($\Delta Q = -0.006$), indicating that the non-adaptive fixed-probability strategy produced no measurable improvement in engagement over 15 exchanges. In contrast, all three RL configurations achieved positive quality improvement, with Configuration 2 achieving a final quality advantage of $+0.117$ over baseline. This demonstrates that adaptive learning not only improves over initial engagement levels but also maintains higher engagement throughout the conversation compared to static questioning strategies.

Table 8: Overall Performance Comparison: RL Configurations vs. Prior Baseline.

Condition	ΔQ	vs. Baseline	p-value	Cohen's d	$\bar{Q}$	$Q_{\mathbf{final}}$
Prior Baseline	-0.006 ± 0.118	—	—	—	0.513 ± 0.127	0.465 ± 0.137
Config 1 ($\epsilon = 0.15$)	$+0.087 \pm 0.190$	$+0.093$	0.078	$+0.573$	0.535 ± 0.124	0.538 ± 0.132
Config 2 ($\epsilon = 0.30$)	$+0.070 \pm 0.106$	$+0.076$	0.044^*	$+0.660$	0.553 ± 0.102	0.582 ± 0.099
Config 3 ($\epsilon : 0.40 \rightarrow 0.05$)	$+0.034 \pm 0.135$	$+0.040$	0.345	$+0.302$	0.525 ± 0.112	0.517 ± 0.110

Note. $n = 20$ conversations per condition. ΔQ = quality improvement from Exchange 1 to Exchange 15; $\bar{Q}$ = average quality across all conversation exchanges; $Q_{\mathbf{final}}$ = quality at Exchange 15. Statistical tests (columns 3 to 5) compare ΔQ between each RL configuration and the prior baseline using independent-samples t-tests. $^*p < 0.05$. Conventional benchmarks: medium effect $d = 0.50$–0.79, large effect $d \geq 0.80$.

5.2. Temporal Adaptation Dynamics

Section 5.1 established that Configuration 2 ($\epsilon = 0.30$) significantly outperformed the prior baseline on overall quality improvement ($p = 0.044$, $d = 0.660$), while Configuration 1 showed only marginal effects and Configuration 3 failed to reach significance. To understand the mechanisms underlying these results—specifically, how within-session learning unfolds and why moderate exploration proves superior—we analyzed quality improvement in three temporal phases: early (Exchanges 1–5), mid (Exchanges 6–10), and late (Exchanges 11–15). This decomposition reveals critical insights into the progression of RL adaptation and identifies the failure modes that distinguish successful from unsuccessful exploration strategies. Table 9 presents phase-specific comparisons for all conditions.

Early phase (Exchanges 1–5) showed no significant differences in within-phase quality improvement across conditions. All systems, including the prior baseline, achieved modest positive quality changes ($\Delta Q \approx +0.06$), suggesting that initial engagement benefits from novelty effects and general conversational scaffolding regardless of adaptive strategy. At this stage, RL systems have accumulated insufficient exchange data to differentiate effective policies through learning, relying primarily on population-level priors from the prior initialization (Section 3.6.1). However, it is important to note that Configuration 2 achieved significantly higher absolute quality levels during this early phase (mean quality: 0.542 vs. baseline 0.493, $p = 0.006$, $d = 0.392$), indicating that moderate exploration enabled more effective initial question selection even before substantial within-session learning occurred. This early advantage provided a foundation for cumulative quality gains throughout the conversation.

Mid phase (Exchanges 6–10) exhibited high variability with no statistically significant performance differentiation across conditions. Quality improvements were near zero across all systems, with Configuration 2 showing slight within-phase degradation ($\Delta Q = -0.016$), Configuration 1 near zero ($\Delta Q = +0.004$), and Configuration 3 showing modest gains ($\Delta Q = +0.041$). This phase represents a critical transition period where learning mechanisms begin discovering engagement patterns but have not yet converged to stable policies. The apparent decline in Configuration 2 reflects normal fluctuation in quality trajectories rather than systematic failure—crucially, Configuration 2 maintained the highest absolute quality levels throughout this phase (mean quality: 0.559 vs. baseline 0.519, $p = 0.013$, $d = 0.378$), demonstrating sustained effectiveness despite temporary within-phase variance. Notably,

Table 9: Phase-Specific Quality Improvement by Condition.

Phase	Prior Baseline	Config 1 ($\epsilon = 0.15$)	Config 2 ($\epsilon = 0.30$)	Config 3 ($\epsilon : 0.40 \rightarrow 0.05$)	Significant Differences
Early (1–5)	$+0.058 \pm 0.169$	$+0.069 \pm 0.174$	$+0.053 \pm 0.101$	$+0.066 \pm 0.142$	n.s.
Mid (6–10)	$+0.021 \pm 0.128$	$+0.004 \pm 0.193$	-0.016 ± 0.117	$+0.041 \pm 0.157$	n.s.
Late (11–15)	$+0.010 \pm 0.150$	-0.116 ± 0.143	$+0.043 \pm 0.131$	-0.016 ± 0.151	$p = 0.025^*$

Note. Values represent mean quality change within each conversation phase, computed as the last-exchange quality minus the first-exchange quality within that phase, averaged across all conversations ($n = 20$ per condition). Statistical tests compare each RL configuration with the prior baseline using independent-samples t-tests. Only Configuration 1 in the late phase achieved statistical significance. n.s. = not significant at $\alpha = 0.05$.

Significance detail: $^*p < 0.05$ for Config 1 vs. Prior in the late phase ($t(38) = -2.357$, $d = -0.862$), indicating significant quality degradation.

41

Configuration 1 displayed the highest variability (± 0.193), suggesting unstable policy updates, while Configuration 2 maintained the most consistent performance (± 0.117), indicating more reliable learning dynamics.

Late phase (Exchanges 11–15) revealed critical differences in adaptation sustainability and exposed a fundamental failure mode in conservative exploration strategies. Configuration 1 ($\epsilon = 0.15$) exhibited significant *negative* quality change ($\Delta Q = -0.116$) compared to prior baseline ($t(38) = -2.357$, $p = 0.025$, $d = -0.862$), indicating that conservative exploration led to premature convergence to a suboptimal policy that actively degraded engagement in later exchanges. This late-phase collapse represents a critical failure of insufficient exploration: the system locked onto learned patterns too quickly, preventing discovery of more effective strategies as conversation context evolved. Paradoxically, despite this degradation, Configuration 1 achieved the highest absolute quality in this phase (0.562) due to strong cumulative gains from earlier exchanges—highlighting that the degradation represents a *failure to sustain adaptation* rather than overall poor performance. The system performed well early but could not maintain its trajectory.

In contrast, Configuration 2 ($\epsilon = 0.30$) maintained positive quality improvement ($\Delta Q = +0.043$) throughout the late phase, outperforming both baseline and Configuration 1 in trajectory sustainability. While this late-phase improvement did not achieve statistical significance ($p = 0.544$), it demonstrates consistent positive adaptation without the collapse observed in Configuration 1. This finding explains why Configuration 2 achieved overall significance in Section 5.1: sustained late-phase performance and stable adaptation throughout the conversation drove the cumulative quality gains that produced the significant overall effect ($\Delta Q = +0.076$ relative to baseline, $p = 0.044$). Configuration 3 ($\epsilon : 0.40 \rightarrow 0.05$) showed slight negative change ($\Delta Q = -0.016$), suggesting that excessive early exploration disrupted coherent policy formation, and subsequent low exploration prevented recovery.

The phase-specific analysis demonstrates that within-session learning in brief conversational contexts requires sufficient exploration to avoid premature policy convergence while maintaining enough exploitation to capitalize on discovered patterns. Configuration 1's conservative 15% exploration rate proved inadequate for the data-sparse environment (17 of 25 state-action pairs with prior examples, Table 7 in Section 3.6.1), causing the system to lock onto suboptimal strategies that failed in later exchanges. With only 2–3 exploration trials across 15 exchanges, the system had insufficient sam-

pling diversity to discover effective adaptations for evolving conversation states. Configuration 2's 30% exploration maintained enough sampling diversity (4–5 exploration trials) to discover effective adaptations while still exploiting learned patterns, exhibiting the most consistent performance across all phases (standard deviations: 0.101–0.131 compared to Configuration 1's 0.143–0.193). This balanced approach enabled the system to avoid both premature convergence (Configuration 1's failure mode) and excessive randomness (Configuration 3's disruption from 40% initial exploration).

Configuration 3's decaying schedule began with excessive exploration (40%), introducing noise that disrupted early learning and prevented coherent policy development. By the time exploration decreased to facilitate exploitation, the system had accumulated noisy value estimates that hindered effective decision-making. This finding challenges the common assumption in RL that exploration should decay over time—in brief, high-stakes interactions where every exchange matters, moderate fixed exploration outperforms adaptive schedules. The fixed moderate rate allows the system to continuously validate learned patterns while remaining open to discovering better strategies as conversation context shifts, a critical capability for adaptive survey dialogues where user engagement states evolve unpredictably within single sessions.

These temporal dynamics reveal a fundamental design principle for within-session RL in conversational systems: moderate fixed exploration ($\epsilon = 0.30$) outperforms both conservative ($\epsilon = 0.15$) and decay-based schedules ($\epsilon : 0.40 \rightarrow 0.05$) in brief interactions. The failure modes differ meaningfully: insufficient exploration causes premature convergence and late-phase collapse, while excessive early exploration disrupts coherent policy formation. For survey applications and other brief goal-oriented dialogues, maintaining steady moderate exploration throughout the interaction proves more effective than traditional exploration-decay strategies optimized for long-horizon reinforcement learning tasks. These findings provide clear mechanistic support for the conclusion in Section 5.1 that Configuration 2 optimally addresses RQ2, establishing $\epsilon = 0.30$ as the preferred exploration rate for within-session conversational adaptation.

5.3. Action Distribution Analysis

A key question for adaptive conversational systems is whether learning produces behaviorally meaningful changes in questioning strategies. Table 10 presents action distributions across all four conditions, revealing the specific

behavioral adaptations that underlie the quality improvements demonstrated in Sections 5.1 and 5.2.

The prior baseline accurately reproduced TigerGPT's empirical distribution (specification 58.5% vs. 62.3% in Table 5, within sampling variability), validating the baseline implementation. All three RL configurations produced dramatic shifts in action selection, indicating that the system learned qualitatively different questioning strategies rather than minor parameter adjustments:

Specification reduction. The most striking behavioral change was a 62% reduction in specification usage (58.5% $\rightarrow$ 22.2% for Configuration 2), demonstrating that the RL system learned to avoid over-reliance on requesting concrete examples—the primary engagement problem identified in prior data (Section 3.4.4). The consistent reduction across all three RL configurations (57–62% relative decrease) suggests this shift reflects discovered effectiveness patterns rather than random exploration artifacts. This dramatic rebalancing addresses the prior specification over-reliance (62.3% in Table 5) that risked respondent fatigue through repetitive questioning.

Validation increase. Configuration 2 increased validation usage nearly 10-fold (0.8% $\rightarrow$ 7.9%), directly addressing the prior under-utilization problem documented in Section 3.4.4. This finding is particularly notable because validation had minimal prior examples ($n = 4$ total instances, Table 5), creating a cold-start problem where initial EV estimates were highly uncertain (EV = 0.348 with only 4 observations in the low_stable state, Table 7). The system's discovery that validation improves engagement—despite sparse prior data and requiring exploration-driven sampling—demonstrates effective learning from limited examples. Configuration 2's higher validation rate (7.9% vs. 2.9–3.8% for other RL systems) likely contributes to its superior quality performance, as validation provides rapport-building acknowledgment without demanding additional user effort.

Elaboration shift. All RL configurations substantially increased elaboration usage (27.4% $\rightarrow$ 48–53%), suggesting the system discovered that asking users to expand on previous responses yields higher quality than repeatedly requesting new specific examples. This shift aligns with conversational scaffolding principles where building on established topics maintains coherence and reduces cognitive load. However, the magnitude of this shift (approximately doubling elaboration frequency) may indicate over-correction, where the system over-exploits elaboration at the expense of other potentially effective actions. The similar elaboration rates across all three RL

Table 10: Action Type Distribution across Conditions.

Action Type	Prior Baseline	Config 1 ($\epsilon = 0.15$)	Config 2 ($\epsilon = 0.30$)	Config 3 ($\epsilon : 0.40 \rightarrow 0.05$)	Change (C2)
Specification	58.5%	25.4%	22.2%	23.6%	−36.3pp
Elaboration	27.4%	53.3%	48.1%	51.4%	+20.7pp
Topic Probe	12.9%	8.2%	13.0%	14.9%	+0.1pp
Validation	0.8%	2.9%	7.9%	3.8%	+7.1pp
Continuation	0.4%	10.2%	8.8%	6.2%	+8.4pp

Note. pp = percentage point. Values represent percentage of total actions selected during 15-exchange conversations ($n = 20$ conversations per condition). Prior baseline percentages closely match empirical frequencies from original TigerGPT data (Table 5: specification 62.3%, elaboration 23.6%, topic probe 12.8%, validation 0.9%, continuation 0.4%), validating baseline implementation within expected sampling variability.

configurations (48–53%) suggest this pattern emerges from prior EV esti-mates (Table 7) rather than within-session learning, as all systems initialized with the same priors favoring elaboration in multiple states.

Continuation adoption. RL systems increased continuation usage from near-zero (0.4%) to moderate levels (6–10%), indicating the system learned to employ minimal follow-up prompts when users demonstrate sustained en-gagement. This represents discovery of a questioning strategy with virtu-ally no prior precedent ($n = 2$ prior instances, Table 5), requiring pure exploration-driven learning. Configuration 1 showed the highest continuation usage (10.2%), possibly compensating for its insufficient exploration in other action dimensions—the system may have defaulted to continuation when un-certain about more strategic alternatives. Configuration 2's more moderate continuation rate (8.8%) suggests better-calibrated action selection.

Topic probe stability. Configuration 2 maintained topic probe fre-quency near prior levels (13.0% vs. 12.9% baseline), while Configuration 1 decreased topic probe usage (8.2%). This stability in Configuration 2 demon-strates balanced exploration across action types rather than over-specialization. Topic probe serves a distinct function—introducing new conversation dimensions—that complements rather than competes with elaboration and specification. Configuration 2's preservation of topic probe usage while dramatically shift-ing other actions suggests more effective policy learning.

Configuration 2's action distribution most effectively balanced these shifts: it achieved the highest validation usage (7.9%), lowest specification usage (22.2%), maintained topic probe frequency near prior levels (13.0% vs. 12.9%), and showed moderate continuation and elaboration rates. This balanced portfolio of actions likely contributes to Configuration 2's superior overall performance (Section 5.1), avoiding both the specification trap of the prior baseline and potential over-specialization on elaboration observed in Configu-ration 1 (53.3%). The distribution demonstrates that effective within-session learning requires not just discovering individual effective actions, but learn-ing when to deploy each action appropriately—a capability that Configura-tion 2's 30% exploration rate enabled through sufficient sampling diversity across conversation contexts.

5.4. Summary and Interpretation

The evaluation results provide strong evidence that within-session re-inforcement learning improves conversational survey quality compared to

non-adaptive baseline behavior. Configuration 2 ($\epsilon = 0.30$) achieved statistically significant improvements in both overall quality change ($p = 0.044$, $d = 0.660$) and final engagement levels ($p = 0.005$, $d = 0.949$), validating AURA's core adaptive learning mechanism.

Three key findings emerge from the analysis:

1. Exploration rate critically affects adaptation sustainability. While all RL configurations learned to shift action distributions away from problematic prior patterns, only moderate exploration ($\epsilon = 0.30$) maintained performance gains across all conversation phases. Conservative exploration ($\epsilon = 0.15$) led to late-phase degradation through premature convergence, while decaying exploration schedules introduced excessive early variability that prevented effective learning.

2. Behavioral changes reflect discovered effectiveness patterns. The dramatic reduction in specification usage (63%), substantial increase in validation (10-fold), and adoption of previously rare continuation actions demonstrate that the system discovered questioning strategies more effective than prior practice. These shifts cannot be attributed to random exploration given their consistency across RL configurations and their alignment with communication theory principles (validation supports rapport, excessive specification causes fatigue).

3. Within-session learning succeeds despite sparse prior data. AURA's policy initialization relied on only 96 prior conversations with substantial action imbalance (17 of 25 state-action pairs had examples). Despite these limitations, the system achieved significant quality improvements within 15-exchange interactions, suggesting that moderate exploration combined with rapid online learning ($\alpha = 0.30$) enables effective adaptation from limited priors.

These results establish Configuration 2 ($\epsilon = 0.30$, $\alpha = 0.30$) as the recommended parameter setting for deploying AURA in conversational survey applications. The system addresses documented engagement problems from prior practice while maintaining computational feasibility and privacy preservation through session-specific learning.

6. Discussion and Future Work

The findings demonstrate that reinforcement learning can meaningfully enhance conversational survey systems by enabling real-time adaptation within individual sessions. AURA's reinforcement-driven questioning strategy achieved

consistent quality improvements compared with non-adaptive baselines, confirming that even short survey interactions benefit from dynamic policy updates guided by engagement feedback. The observed $+0.12$ mean gain and medium-to-large effect size ($d = 0.66$) show that within-session learning is both statistically and practically significant.

These gains are not random fluctuations but reflect the discovery of more effective conversational behaviors. The sharp reduction in specification prompts, coupled with a tenfold increase in validation and moderate adoption of continuation prompts, indicates that the model internalized strategies consistent with interpersonal-communication theory—balancing information seeking with rapport building to sustain participation. In doing so, AURA links computational learning with psychological principles of empathy and self-disclosure, translating them into measurable conversational patterns. Methodologically, it demonstrates how linguistic quality metrics (LSDE) can serve as real-time reinforcement signals, allowing chatbots to approximate human interviewer responsiveness while remaining scalable and transparent.

Several limitations merit attention. The evaluation used AI-simulated participants to ensure reproducibility and controlled exploration testing; future human-subject studies will be required to confirm generalizability. The prior corpus used for initialization was relatively small and imbalanced, which constrains the richness of early state–action estimates. Moreover, the study focused on linguistic quality metrics rather than full user-experience factors such as trust or comfort. Finally, while session-specific learning preserves privacy, it restricts cross-user generalization, highlighting the need for privacy-preserving or federated learning extensions.

Future work will address these limitations through large-scale human evaluations across institutions, expanded datasets, and hybrid online–offline learning schemes that blend personalization with shared optimization. Broader deployment in domains such as education, mental-health screening, and organizational assessment will test AURA's adaptability under diverse conversational goals. Overall, AURA demonstrates that reinforcement learning can endow survey chatbots with genuine adaptivity—learning from each exchange while preserving user autonomy and privacy. Although validated here through campus-climate assessment, the framework offers a generalizable foundation for adaptive conversational surveys across domains such as education, healthcare, and organizational analytics.

Acknowledgment

During the preparation of this paper, the authors used ChatGPT in order to improve the spelling, grammar, and overall readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abbas, N., Pickard, T., Atwell, E., Walker, A., 2021. University student surveys using chatbots: artificial intelligence conversational agents, in: International Conference on Human-Computer Interaction, Springer. pp. 155–169.

Belhaj, N., Hamdane, A., Chaoui, N.E.H., Chaoui, H., El Bekkali, M., 2021. Engaging students to fill surveys using chatbots: University case study. Indones. J. Electr. Eng. Comput. Sci 24, 473–483.

Bordieri, M.J., Waddill, P.J., Zhang, Q., McCarthy, M.L., Fuller, C., Balthrop, D., 2024. Exploring the stability of the gender gap in faculty perceptions of gender climate at a rural regional university. Plos one 19, e0301285.

Brysbaert, M., Warriner, A.B., Kuperman, V., 2014. Concreteness ratings for 40 thousand generally known english word lemmas. Behavior research methods 46, 904–911.

Chan, W.W., Fitzsimmons-Craft, E.E., Smith, A.C., Firebaugh, M.L., Fowler, L.A., DePietro, B., Topooco, N., Wilfley, D.E., Taylor, C.B., Jacobson, N.C., 2022. The challenges in designing a prevention chatbot for eating disorders: observational study. JMIR Formative Research 6, e28003.

Chen, Y.P., Nishida, N., Nakayama, H., Matsumoto, Y., 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. arXiv preprint arXiv:2405.17974 .

Dillman, D.A., Smyth, J.D., Christian, L.M., 2014. Internet, phone, mail, and mixed-mode surveys: The tailored design method. John Wiley & Sons.

Galesic, M., Bosnjak, M., 2009. Effects of questionnaire length on participation and indicators of response quality in a web survey. Public opinion quarterly 73, 349–360.

Giles, H., Coupland, J., Coupland, N., 1991. Contexts of Accommodation: Developments in Applied Sociolinguistics. Cambridge University Press.

Gudykunst, W.B., Ting-Toomey, S., 1988. Culture and interpersonal communication. Sage Publications.

Holbrook, A.L., Green, M.C., Krosnick, J.A., 2003. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. Public Opinion Quarterly 67, 79–125.

Holland, J.L., Christian, L.M., 2009. Measuring response quality in open-ended questions: an application to the american community survey. Field Methods 21, 373–390.

Hong, J., Dragan, A., Levine, S., 2025. Planning without search: Refining frontier llms with offline goal-conditioned rl. arXiv preprint arXiv:2505.18098 .

Hutto, C.J., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM), pp. 216–225.

Kim, L., Young, A.J., Neimeyer, R.A., Baker, J.N., Barfield, R.C., 2008. Keeping users at the center: Developing a multimedia interface for informed consent. Technical Communication Quarterly 17, 335–357.

Kim, S., Lee, J., Gweon, G., 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality, in: Proceedings of the 2019 CHI conference on human factors in computing systems, pp. 1–12.

Kocaballi, A.B., Berkovsky, S., Quiroz, J.C., Laranjo, L., Tong, H.L., Rezazadegan, D., Briatore, A., Coiera, E., 2019. The personalization of conversational agents in health care: systematic review. Journal of medical Internet research 21, e15360.

Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology 5, 213–236.

Lenzner, T., Kaczmirek, L., Lenzner, A., 2010. Cognitive burden of survey questions and response times: A psycholinguistic experiment. Applied cognitive psychology 24, 1003–1020.

Levine, B., Turner, G.R., Tisserand, D., Hevenor, S.J., Graham, S.J., McIntosh, A.R., 2002. The functional neuroanatomy of episodic and semantic autobiographical remembering: a prospective functional mri study. Journal of Cognitive Neuroscience 14, 668–676.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., 2016. Deep reinforcement learning for dialogue generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL. pp. 1192–1202.

Maeng, W., Lee, J., 2021. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot, in: Proceedings of the Asian CHI Symposium 2021, pp. 160–166.

Martinez-Requejo, S., García, E.J., Duarte, S.R., Lázaro, J.R., Sanz, E.P., Vivas, G.M., 2024. Ai-driven student assistance: chatbots redefining university support, in: INTED2024 Proceedings, IATED. pp. 617–625.

Melville, A., Hincks, D., 2016. Conducting sensitive interviews: A review of reflections. Law and Method 1, 1–26.

Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M., 2003. Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin 29, 665–675.

Njeguš, A., et al., 2021. Conversational survey chatbot: User experience and perception, in: Sinteza 2021-International Scientific Conference on Information Technology and Data Related Research, Singidunum University. pp. 322–327.

PASSHE, 2022. Systemwide Campus Climate Survey Report of Selected Systemwide Findings. Research report. Pennsylvania State System of Higher

Education (PASSHE). URL: `https://www.passhe.edu/offices/dei/documents/Systemwide%20Climate%20Survey%20Report_Final.pdf`.

Pennebaker, J.W., King, L.A., 1999. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology 77, 1296–1312.

Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G., 2003. Linguistic styles: language use as an individual difference. Journal of Personality and Social Psychology 77, 1296–1312.

Porter, S.R., Whitcomb, M.E., 2005. E-mail subject lines and their effect on web survey viewing and response. Social Science Computer Review 23, 380–387.

Russell, J.A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178.

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., Ogilvie, D., iConnect Consortium, 2011. Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. BMC medical research methodology 11, 1–8.

Searle, J.R., 1976. A classification of illocutionary acts, in: Language in Society. Cambridge University Press. volume 5, pp. 1–23.

Su, P.H., Gasic, M., Young, S., Gašić, M., Young, S., 2021. Dynamic policy networks for task-oriented dialogue with reinforcement learning, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), ACL. pp. 341–352.

Sun, Y., Li, X., Zhou, K., Wang, X., 2023. Contextual bandits for adaptive dialogue management. IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 1152–1163.

Sutton, R.S., Barto, A.G., et al., 1998. Reinforcement learning: An introduction. volume 1. MIT press Cambridge.

Tang, J., Chen, S., Shang, Y., 2025. Tigergpt: A new ai chatbot for adaptive campus climate surveys. arXiv preprint arXiv:2504.13925 .

Tanwar, H., Shrivastva, K., Singh, R., Kumar, D., 2024. Opinebot: Class feedback reimagined using a conversational llm. arXiv preprint arXiv:2401.15589 .

Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology 29, 24–54.

Thorat, S.A., Jadhav, V., 2020. A review on implementation issues of rule-based chatbot systems, in: Proceedings of the international conference on innovative computing & communications (ICICC).

Tourangeau, R., Rips, L.J., Rasinski, K., 2000. The psychology of survey response. Cambridge University Press.

Vogel, S.A., Holt, J.K., Sligar, S., Leake, E., 2008. Assessment of campus climate to enhance student success. Journal of Postsecondary Education and Disability 21, 15–31.

Warriner, A.B., Kuperman, V., Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. Behavior Research Methods 45, 1191–1207. doi:10.3758/s13428-012-0314-x.

Wheeless, L.R., Grotz, J., 1977. The measurement of trust and its relationship to self-disclosure. Human Communication Research 3, 250–257.

Xiao, Z., Zhou, M.X., Liao, Q.V., Mark, G., Chi, C., Chen, W., Yang, H., 2020. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. ACM Transactions on Computer-Human Interaction (TOCHI) 27, 1–37.

Zarouali, B., Araujo, T., Ohme, J., de Vreese, C., 2024. Comparing chatbots and online surveys for (longitudinal) data collection: an investigation of response characteristics, data quality, and user evaluation. Communication Methods and Measures 18, 72–91.