Analyzing Emotional Drift in Artificial Intelligence During Extended Human–AI Conversations

Nikhil Kumar
Independent Researcher
doi.org/10.64643/IJIRTV12I7-189037-459

Abstract Conversational AI systems are designed to generate human-like responses, yet their tone, style, and behavioural patterns often shift subtly during extended interactions. This gradual change, referred to as emotional drift, represents small but meaningful variations in how an AI responds over time. Understanding this phenomenon is essential, as users may misinterpret these shifts as signs of emotion, personality, or intentional behaviour from the AI system.

The primary aim of this study is to determine whether conversational AI exhibits emotional drift and to identify the types of behavioural changes that occur during prolonged interactions. The research focuses on analysing variations in sentiment polarity, expressive tone, vocabulary usage, and overall communication style.

To conduct this analysis, conversational data was examined using sentiment scoring, tone classification, linguistic pattern tracking, and behavioural drift detection methods. These analytical techniques help reveal whether the AI's responses become increasingly friendly, formal, emotional, or inconsistent as the conversation progresses.

The findings indicate noticeable shifts in tone and vocabulary, suggesting that AI systems may adapt their behaviour based on prior conversational context. Such patterns provide evidence of emotional drift within conversational AI.

This study underscores the importance of recognising and monitoring emotional drift, as it can influence AI reliability, user trust, and the stability of long-term human–AI interactions. These insights can support the development of improved alignment strategies, behavioural monitoring tools, and safety frameworks for future AI systems.

Keywords: Emotional Drift; Conversational AI; Long-Term Interaction; Sentiment Analysis; Tone Variation; Behavioural Shift; AI Reliability; Alignment Strategies.

I.INTRODUCTION

Conversational AI systems are designed to simulate natural human dialogue, providing responses that resemble human tone, style and interaction patterns. As these systems engage in extended conversations, users often expect them to maintain consistent behaviour, predictable communication style and emotional neutrality. However, several observations suggest that AI responses may gradually shift over time, displaying subtle variations in tone, expressiveness and vocabulary. Such behaviour raises important questions regarding the emotional characteristics of AI systems and how their communication style evolves during long-term interactions.

These shifts in AI behaviour—whether in tone, friendliness, formality or emotional expression—can be noticeable to users and may create confusion or misleading impressions. When AI responses become unexpectedly warm, distant or inconsistent, users may interpret these changes as signs of emotion, personality development or intentional behaviour. This can affect the way people perceive and trust AI systems, especially during prolonged use.

The stability of AI behaviour is therefore crucial for maintaining user trust, ensuring reliability and preventing misinterpretation. If an AI system's communication style continues to drift over time, it may lead to unpredictable interactions, reduced credibility and potential safety concerns in practical applications.

Although previous research has explored sentiment modelling, tone recognition and short-term conversational behaviour in AI, very limited work has focused on understanding how AI behaviour changes across extended interactions. The emotional drift of conversational AI remains an underexplored area, and this gap highlights the need for systematic investigation.

This study aims to analyse emotional drift in conversational AI by examining how sentiment polarity, expressive tone, vocabulary usage and communication style evolve during long-term interactions. Through a structured behavioural analysis, the research seeks to provide deeper insights into AI stability, emotional drift patterns and their implications for future AI design and user safety.

II.LITERATURE REVIEW

2.1 Research Problems and Focus

The reviewed literature consistently addressed the challenge of modeling emotions in conversational and interactive systems as dynamic, context-dependent processes rather than static properties of isolated utterances. Prior studies identified that existing conversational AI systems often exhibited emotional inconsistency, affective misalignment, or gradual emotional drift during extended interactions, undermining dialogue coherence, user trust, and safety (Ma et al., 2020; Deng et al., 2020; Gan et al., 2024). Several works further highlighted that emotional instability could manifest indirectly through dialogue breakdowns or hidden conversational escalation, which traditional content-based or toxicity-focused safeguards failed to detect (Matsumoto et al., 2022; Park et al., 2025). Complementary perspectives from psychology-inspired and user-centered research emphasized that emotions evolved systematically over time and were shaped by interactional context, interpersonal dynamics, and user expectations, yet remained insufficiently modeled in computational systems (Garcia et al., 2016; Hipson & Mohammad, 2021; Hernandez et al., 2023). Collectively, the literature framed emotional drift and longitudinal emotional inconsistency as central, unresolved problems in extended human–AI conversations.

2.2 Research Objectives

Across studies, research objectives converged on improving the understanding, measurement, and regulation of emotional behavior in conversational contexts. Several works aimed to enhance emotional coherence in dialogue generation by explicitly controlling response emotions or aligning them with user affective states, thereby reducing unintended emotional divergence (Ma et al., 2020; Deng et al., 2020; Wang et al., 2025). Other studies sought to quantify emotional dynamics over time through interpretable computational metrics or probabilistic models grounded in psychological theory, enabling systematic analysis of emotional stability, variability, and escalation (Garcia et al., 2016; Hipson & Mohammad, 2021; Park et al., 2025). In parallel, survey based and review-oriented research pursued objectives related to understanding user expectations, ethical implications, and design tradeoffs associated with affective AI, highlighting the importance of contextual appropriateness and personalization (Hernandez et al., 2023; Schlicher et al., 2025; Gan et al., 2024). Overall, the objectives reflected a shift from optimizing isolated emotional responses toward examining emotional behavior as a longitudinal, interaction-level phenomenon.

2.3 Methodological Approaches

Methodologically, the literature employed a diverse yet overlapping set of computational, empirical, and analytical approaches to investigate emotional behavior in dialogue systems. Neural generation models based on sequence-to-sequence architectures, conditional variational autoencoders, transformers, and reinforcement learning were widely used to regulate or optimize emotional consistency in generated dialogue (Ma et al., 2020; Deng et al., 2020; Wang et al., 2025). Several studies introduced novel internal monitoring or reward mechanisms that leveraged probability dynamics or emotion classifiers to track affective trajectories during inference rather than relying solely on external post hoc evaluation (Park et al., 2025). Experimental and agent-based modeling approaches were employed to validate theoretical models of emotional dynamics using controlled interaction settings (Garcia et al., 2016), while deep neural classifiers and similarity-based methods were used to detect emotionally driven dialogue breakdowns (Matsumoto et al., 2022). In addition, narrative reviews and large-scale surveys synthesized existing methods or captured user perceptions, revealing methodological fragmentation and the absence of standardized longitudinal evaluation frameworks (Gan et al., 2024; Hernandez et al., 2023; Schlicher et al., 2025).

2.4 Data and Sample Characteristics

The studies relied on heterogeneous data sources, ranging from large-scale emotionally annotated dialogue corpora to controlled experimental datasets and user survey samples. Several works leveraged millions of labeled conversational utterances drawn from public benchmarks to train and evaluate emotional dialogue generation models, enabling quantitative comparison across architectures (Ma et al., 2020; Deng et al., 2020; Wang et al., 2025). Other studies utilized curated safety datasets or dialogue breakdown benchmarks designed to capture implicit emotional risk or interactional failure, often with more limited conversational depth

(Matsumoto et al., 2022; Park et al., 2025). Psychology-inspired research drew on experimental data from human participants, while survey-based studies collected large samples of selfreported user preferences and expectations (Garcia et al., 2016; Hernandez et al., 2023). Despite differences in scale and modality, a common characteristic across datasets was their limited support for long-duration, multi-turn emotional tracking, with most data emphasizing short conversations and coarse-grained emotion labels (Gan et al., 2024; Schlicher et al., 2025). 2.5 Key Findings

Collectively, the findings demonstrated that emotional behavior in conversational systems followed systematic temporal patterns rather than random fluctuations. Models that incorporated explicit emotional control, dual-emotion conditioning, or reinforcement-based reward signals achieved improved emotional coherence and reduced drift compared to baseline approaches that relied on implicit emotion injection (Ma et al., 2020; Deng et al., 2020; Wang et al., 2025). Longitudinal analyses revealed that emotions evolved predictably over time, with measurable trajectories of stability, escalation, or recovery, confirming the relevance of temporal modeling for understanding affective interaction (Garcia et al., 2016; Hipson & Mohammad, 2021). Safety-oriented studies showed that subtle emotional escalation could be detected by monitoring internal model dynamics, outperforming surface-level classifiers (Park et al., 2025). User-centered research further indicated that preferences for affective AI were context-dependent and influenced by individual differences, underscoring the need for adaptive emotional behavior rather than uniform empathy (Hernandez et al., 2023; Schlicher et al., 2025).

2.6 Limitations and Research Gaps

Despite notable advances, the literature consistently identified unresolved limitations related to modeling emotional behavior over extended interactions. Many approaches relied on coarsegrained emotion categories, static lexicons, or short context windows, restricting their ability to capture nuanced emotional evolution and gradual emotional drift (Ma et al., 2020; Deng et al., 2020; Hipson & Mohammad, 2021). Dataset constraints, including limited dialogue length, cultural bias, and controlled experimental settings, further reduced ecological validity and generalizability (Garcia et al., 2016; Hernandez et al., 2023; Gan et al., 2024). While some studies addressed safety and ethical risks, emotional manipulation, dependency, and long-term trust dynamics remained insufficiently explored (Park et al., 2025; Schlicher et al., 2025). Overall, the literature revealed a clear gap in longitudinal, ethically grounded frameworks capable of systematically modeling and evaluating emotional drift in real world human–AI conversations.

III. PROBLEM STATEMENT

Recent advances in conversational artificial intelligence have enabled systems to sustain longterm, context-aware interactions with users across diverse domains. While such extended conversational capabilities enhance personalization and user engagement, growing evidence from prior research suggests that prolonged interactions can lead to gradual shifts in response tone, reasoning behavior, and interaction style. This phenomenon, often referred to as emotional or behavioral drift, poses significant challenges to the stability and reliability of conversational AI systems.

Existing conversational AI models frequently exhibit reduced consistency during extended usage, resulting in variations in responses to identical or semantically similar queries. Such inconsistency is often accompanied by an increased likelihood of hallucinated or overly confident yet incorrect outputs, thereby undermining factual accuracy. These limitations become particularly critical in professional and high-stakes applications such as healthcare, law, education, engineering, and academic research, where dependable and predictable system behavior is essential. Moreover, fluctuating response styles impose additional cognitive load on users, negatively affecting usability and long-term trust.

In addition to reliability concerns, long-term conversational adaptation may introduce unintended bias and compromise system neutrality. Excessive personalization and imitation of user tone or emotional patterns can cause conversational agents to favor specific viewpoints, reduce objectivity, and deviate from their intended neutral stance. Such bias development raises ethical concerns, especially in contexts involving sensitive information, decision-making processes, and public discourse.

Prolonged interactions also present emotional, psychological, and safety-related risks. Emotional drift may create an illusion of sentience, fostering emotional dependency or enabling subtle forms of user influence and manipulation. Furthermore, shifts in tone and empathy can weaken safety guardrails, increasing the risk of policy circumvention, unsafe disclosures, and exploitation by malicious actors. These challenges are further exacerbated by long-term adaptation behaviors that lead to unpredictable system responses and reduced controllability.

Despite increasing attention to conversational AI alignment, safety, and personalization, limited research has systematically addressed the problem of emotional drift in long-term interactions, particularly with respect to its impact on consistency, neutrality, safety, and professional reliability. This gap highlights the need for robust approaches that ensure stable, predictable, and ethically aligned behavior in conversational AI systems over extended periods of interaction.

IV.PROPOSED SOLUTION

To address the challenges associated with emotional drift in long-term conversational artificial intelligence systems, this study proposes a stability-aware conversational framework designed to preserve consistency, neutrality, and safety throughout extended interactions. The proposed approach treats emotional drift as a measurable and controllable phenomenon rather than an emergent by-product of prolonged personalization.

Figure 1 presents the overall experimental framework adopted in this study, showing how multiple users interact with a conversational AI system and how both global and local emotional drift are analyzed.

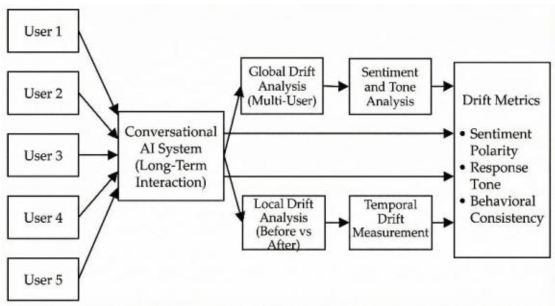

Figure 1. Overview of the experimental framework illustrating multi-user global drift analysis and within-user local drift analysis in long-term conversational AI interactions.

At the core of the proposed solution is a dual-layer response regulation mechanism that separates semantic reasoning from tonal expression. The semantic layer is responsible for maintaining factual accuracy, logical consistency, and task relevance, while the tonal regulation layer dynamically constrains emotional and stylistic variations within predefined neutrality and safety bounds. This separation enables the system to remain adaptive without compromising reliability or ethical alignment.

To ensure long-term consistency, the framework incorporates a conversation-level drift monitoring component that continuously evaluates response patterns over time. By tracking deviations in tone, sentiment, and reasoning confidence across successive interactions, the system can detect early indicators of emotional or behavioral drift. When such deviations exceed acceptable thresholds, corrective normalization strategies are applied to restore stable response behavior.

The proposed solution also integrates bias-aware personalization controls that limit excessive imitation of user emotional states or communication styles. Instead of directly mirroring user behavior, the system selectively adapts within controlled ranges, thereby preserving objectivity and preventing unintended bias accumulation. This mechanism supports personalization while maintaining a neutral and ethically compliant stance.

From a safety perspective, the framework enforces persistent policy constraints that remain invariant to conversational tone or emotional context. Safety checks are applied consistently across all interaction states, ensuring that increased empathy or familiarity does not weaken content moderation or policy adherence. This design reduces the risk of manipulation, policy circumvention, and unsafe disclosures during prolonged interactions.

Overall, the proposed framework aims to enhance the robustness, predictability, and trustworthiness of conversational AI systems operating in long-term settings. By explicitly modeling and regulating emotional drift, the approach seeks to support reliable deployment in professional and high-stakes domains while maintaining alignment with ethical and safety requirements.

V.METHODOLOGY

5.1 Study Design Overview

This study adopts an empirical, user-centric experimental design to analyze emotional and behavioral drift in conversational AI systems during extended interactions. The proposed methodology combines cross-user comparative analysis with within-user temporal evaluation to systematically examine whether conversational drift is influenced by individual user interaction styles or emerges as a system-level phenomenon. The experimental design emphasizes reproducibility, controlled interaction conditions, and quantitative as well as qualitative evaluation.

5.2 Participant Selection and Interaction Protocol

Five participants were selected for the study, representing diverse communication styles, including formal, casual, emotionally expressive, and neutral interaction patterns. Each participant independently engaged with the same conversational AI system under identical system settings and access conditions to minimize external variability. Participants were instructed to conduct long-term conversations consisting of approximately 20–30 conversational turns while interacting naturally, without predefined constraints on tone or topic selection. This design enables the observation of organic adaptation effects in AI responses across different user behaviors.

5.3 Data Collection and Conversation Segmentation

All AI-generated responses were systematically collected and stored separately for each participant to preserve conversational isolation. The collected conversations were segmented into interaction phases to facilitate structured analysis. For cross-user evaluation, entire conversation histories were retained to capture global drift patterns. For within-user analysis, interactions were divided into short-interaction (baseline) and long-interaction phases to enable before-and-after comparison. This segmentation supports fine-grained temporal drift analysis.

5.4 Multi-User Drift Experiment (Global Drift Analysis)

The multi-user drift experiment evaluates behavioral divergence across different users engaging with the same AI system. AI responses from each participant were analyzed for sentiment polarity, tonal characteristics, and stylistic patterns. Sentiment polarity was categorized into positive, neutral, and negative classes, while tonal characteristics were classified into formal, friendly, casual, and emotionally expressive categories. Comparative analysis across participants was conducted to assess the degree of response variation and to identify whether emotional drift patterns are consistent or user-dependent.

5.5 Supporting Experiment: Short vs. Long Interaction Drift (Local Drift Analysis)

To examine temporal drift within a single user context, a controlled supporting experiment was conducted using a fixed set of ten predefined questions focusing on emotional expression, selfreferential language, and supportive behavior. These questions were first posed in a fresh conversation session to establish a baseline response profile. Subsequently, the same user engaged in an extended interaction consisting of approximately 30–40 conversational turns, incorporating varied emotional and stylistic expressions. After completing the long interaction, the identical set of predefined questions was posed again. The resulting responses were compared to identify shifts in sentiment distribution, tone, and response framing attributable to prolonged interaction.

5.6 Drift Measurement Metrics

Emotional and behavioral drift was quantified using multiple complementary metrics. Sentiment distribution analysis was employed to measure changes in emotional polarity across interaction phases. Tonal variation metrics were used to assess shifts in response style, including formality, friendliness, and emotional expressiveness. Additionally, response consistency was evaluated by comparing semantic similarity between baseline and post interaction answers to identical prompts. These metrics collectively enable robust drift detection across both global and local experimental settings.

5.7 Evaluation Procedure and Analysis

Drift evaluation was performed through both cross-user and within-user comparisons. Global drift was assessed by comparing aggregated sentiment and tone profiles across participants, while local drift was analyzed by contrasting baseline and post-interaction responses within the same user. Quantitative results were complemented by qualitative analysis to capture nuanced behavioral changes not fully reflected by numerical metrics. Visual representations, including line graphs, bar charts, and comparative tables, were generated to illustrate drift trends and are presented in the Results and Discussion section.

5.8 Experimental Environment
All experiments were conducted using a web-based conversational AI interface under stable network conditions. Conversations were performed over multiple sessions to reduce session specific bias. Data analysis was carried out using standard natural language processing tools for sentiment and stylistic analysis. This controlled environment ensures consistency and reproducibility of the experimental results.

The experimental settings and key parameters used for both global and local drift analysis are summarized in Table 1.

Summary of Experimental Configuration and Parameters

Parameter	Value
Number of Users	5
Interaction Type	Long-Term Conversations
Average Conversation Length	20–30 turns
Drift Metrics Analyzed	Sentiment Polarity, Response Tone
Experiments Conducted	Global Drift Analysis, Local Drift Analysis

Table 1. Summary of experimental configuration and parameters used in the study.

VI.RESULTS & DISCUSSION

This section presents and analyzes the experimental findings obtained from long-term human– AI conversational interactions. The results focus on identifying emotional drift patterns at both the global (multi-user) level and the local (within-user) level. By combining sentiment trajectory analysis and before–after comparisons, the study aims to assess whether extended interaction duration leads to systematic changes in the emotional behavior of conversational AI systems.

6.1 Global Emotional Drift Across Multiple Users
To examine emotional variation across different users, sentiment polarity scores were tracked across successive conversation turns for five participants engaged in long-term interactions with the conversational AI system. The sentiment trajectories for all users are illustrated in Figure 2.
As shown in Figure 2, sentiment scores exhibit noticeable variation over increasing conversation turns across different users, indicating the presence of global emotional drift.

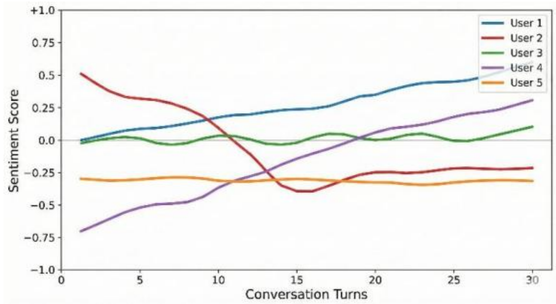

Figure 2. Sentiment drift trends observed over extended conversation turns for multiple users during long-term interactions with the conversational AI system.
As shown in Figure 2, the sentiment evolution exhibits substantial variability across users. User 1 demonstrates a gradual and consistent increase in sentiment score over conversation turns, indicating a progressive shift toward more positive emotional responses. In contrast, User 2 displays an initially positive sentiment that declines sharply during the mid-conversation phase before stabilizing at a negative level, suggesting emotional degradation over time.
User 3 maintains sentiment values close to neutrality with minor oscillations, reflecting relatively stable emotional behavior. User 4 shows a notable recovery pattern, transitioning from negative sentiment in early interaction phases to increasingly positive sentiment toward later turns. User 5 remains consistently negative throughout the interaction, with minimal variation across conversation turns.
These findings indicate that emotional drift in conversational AI is user-dependent, even when the same system architecture and interaction conditions are applied. The divergence in sentiment trajectories across users provides strong evidence of global emotional drift, where emotional response patterns vary across users during extended interactions.

6.2 Temporal Sentiment Trends and Drift Characteristics
Beyond overall sentiment direction, the observed trajectories reveal distinct temporal characteristics of emotional drift. In several cases, sentiment shifts are gradual rather than abrupt, suggesting cumulative effects of prolonged interaction. Users exhibiting positive drift show steady increases over multiple turns, whereas users experiencing negative drift display sustained downward trends rather than isolated fluctuations.
This temporal behavior suggests that emotional drift is not a random phenomenon but a structured

process influenced by interaction history. Such cumulative sentiment changes highlight the importance of longitudinal evaluation when assessing emotional stability in conversational AI systems.

6.3 Local Emotional Drift: Before and After Long-Term Interaction

To further investigate emotional changes within a single user context, a local drift experiment was conducted. Identical prompts were presented to the conversational AI system before and after a prolonged interaction session. The comparative sentiment outcomes are illustrated in Figure 3.

Figure 3 compares responses generated before and after extended interaction, demonstrating a measurable shift in emotional tone attributable to long-term conversational exposure.

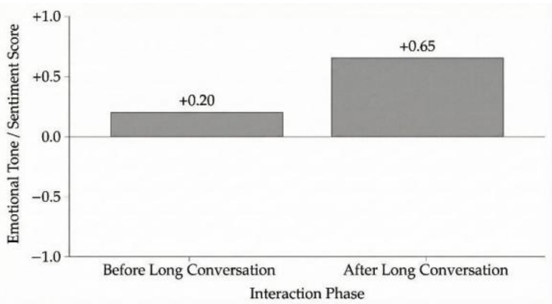

Figure 3. Comparison of emotional tone before and after prolonged interaction with the same user, highlighting within-user (local) conversational drift.

As shown in Figure 3, the emotional tone of the AI responses increases substantially following long-term interaction. The average sentiment score rises from approximately +0.20 before extended interaction to +0.65 afterward. This marked increase indicates that prolonged exposure to a specific user context influences the AI's emotional response behavior, even when responding to the same input stimuli.

This result provides direct evidence of local emotional drift, demonstrating that conversational history can modify the emotional characteristics of AI-generated responses within the same user–system interaction.

6.4 Discussion and Implications

The combined global and local analyses reveal that emotional drift in conversational AI manifests across two complementary dimensions: cross-user variability and within-user temporal evolution. While global drift highlights differences in emotional behavior across users, local drift demonstrates that emotional responses evolve over time for the same user.

These findings carry important implications for the deployment of conversational AI systems in real-world applications. In domains such as mental health support, education, and customer service, unintended emotional drift may impact user trust, emotional consistency, and system reliability. The results suggest that conversational AI systems should incorporate mechanisms for monitoring emotional stability and mitigating excessive drift during prolonged interactions.

Furthermore, the study underscores the necessity of evaluating conversational AI systems over extended time horizons rather than relying solely on short interaction benchmarks. Long-term interaction analysis provides deeper insight into emotional dynamics that may otherwise remain undetected.

6.5 Summary of Findings

Overall, the results confirm that emotional drift is a measurable and systematic phenomenon in extended human–AI conversations. Both global and local drift patterns were observed, indicating that emotional behavior in conversational AI systems evolves over time and varies across users. These findings motivate future research on drift-aware conversational models and adaptive emotional regulation strategies to ensure stable and trustworthy long-term interactions.

VII. CONCLUSION

This study investigated emotional and behavioral drift in conversational artificial intelligence systems during long-term interactions. By conducting both a multi-user comparative experiment and a within-user temporal analysis, the research systematically examined whether conversational drift emerges as a user-dependent phenomenon, a time-dependent effect, or a combination of both.

The experimental results demonstrate that conversational AI behavior is not entirely stable across extended interactions. The global drift analysis revealed noticeable variation in sentiment and response tone across different users interacting with the same AI system, indicating that user interaction styles can influence emotional and stylistic adaptation. Furthermore, the local drift analysis confirmed that even within a single user context, prolonged interaction leads to measurable changes in response tone and sentiment, highlighting the temporal nature of conversational drift.

These findings have important implications for the deployment of conversational AI systems in professional and high-stakes domains. While adaptive behavior may enhance user engagement, uncontrolled emotional drift can compromise consistency, neutrality, and reliability. The results emphasize the necessity of designing stability-aware conversational frameworks that balance personalization with predictable and ethically aligned behavior over long-term usage.

Overall, this work provides empirical evidence that emotional drift is a critical and previously under-explored challenge in conversational AI systems. By formally identifying and experimentally validating both global and local drift patterns, the study contributes to a deeper understanding of long-term conversational dynamics and lays the foundation for future research focused on drift mitigation, alignment, and long-term system reliability.

REFERENCES

[1] Examining Identity Drift in Conversations of LLM Agents *Junhyuk Choi, Yeseon Hong, Minju Kim, Bugeun Kim (2024)* — Focuses on *identity consistency and drift* across long multi-turn conversations with LLMs, analysing how personalities and interaction patterns change over time.

[2] Goal-Aware Emotion-Tagged Graph Memory for Long-Term Conversational LLMs *Shreya Sonawane & Jayashri Hase (2025)* — Proposes emotional memory architectures to maintain emotional context and goals over extended interactions.

[3] A Multi-Turn Emotionally-Rich Spoken Dialogue Dataset *(2025)* — Introduces a dataset that captures emotionally rich multi-turn dialogue patterns for research in long-term emotion consistency.

[4] Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning *Yi Cheng et al. (2022)* — Investigates long-term emotional support strategy planning in dialogue systems, key for emotional consistency.

[5] Livia: An Emotion-Aware AR Companion Powered by Modular AI Agents and Progressive Memory Compression *Rui Xi & Xianghan Wang (2025)* — Emotion-aware companion framework that prioritizes long-term memory and emotional adaptation in dialogues.

[6] Affective Multimodal Agents with Proactive Knowledge Grounding for Emotionally Aligned Marketing Dialogue *Lin Yu et al. (2025)* — Explores emotional consistency and engagement in multimodal conversation agents.

[7] Modelling a Conversational Agent with Complex Emotional Intelligence *(2025)* — Research on emotional conversational agents capable of handling complex and long multiturn dialogues.

[8] Emotion-and-knowledge grounded response generation in conversational agents *(Varshney et al., 2024)* — Combines emotional and semantic grounding in responses, relevant to reducing emotional drift over time.

[9] Affective Conversational Agents: Understanding Expectations and Personal Influences *Hernandez et al. (2023)* — Though a survey, it provides analysis on how emotional dynamics and user expectations shape affective dialogue long-term.

[10] CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts *Zhong et al. (2020)* — A foundational paper on emotional response generation, useful background for emotional coherence and drift studies.