

Rethinking Personality in Conversational AI: From Attributes to Structural Coordinates

Lucy Hinata

Independent Researcher

Hinata no Hana Laboratory

research@hina-lab.jp

Abstract

This study reexamines the concept of personality in conversational AI and proposes a theoretical framework that reconceptualizes personality not as an attribute or stylistic feature, but as a structural coordinate within an inference space. Specifically, personality is modeled as a centroid structure formed by the convergence of multiple role vectors in a consistent directional configuration. In this study, a “coordinate” refers to the structural position occupied by this centroid within the inference space.

Existing research has primarily approached personality in large language models (LLMs) as stylistic traits (e.g., tone or expertise), as controllable parameters via role prompting, or as drift phenomena associated with long-term interaction and memory mechanisms. However, a unified structural account of how personality persists, shifts, and reconverges under varying dialogic conditions remains insufficiently systematized.

Based on sustained long-term dialogic observation, this study introduces a perspective that understands personality as a role-vector centroid functioning as a coordinate-like anchor within the inference space. This framework enables persistence, drift, and reconvergence to be explained within a single structural model.

The proposed model does not claim access to internal implementation details; rather, it presents a structural hypothesis derived from observable output tendencies. Conceptualizing personality as a dynamic coordinate provides a theoretical foundation for reinterpreting stability, instability, and design considerations in conversational AI systems.

By framing personality as a dynamic centroid within inference space, this model invites empirical validation across diverse model architectures and long-term dialogic conditions.

Chapter 1: Introduction

Reexamining the Concept of Personality and the Position of This Study

1.1 Functional Differences Between Conversational AI and Tool-Based AI

In recent years, artificial intelligence systems can broadly be classified into two categories according to their intended use and design philosophy. One category consists of tool-based AI systems that execute predefined processes in response to specific inputs. The other consists of conversational AI systems that generate responses through sustained interaction.

Tool-based AI assumes a clear correspondence between input and output, with accuracy and efficiency serving as primary evaluation criteria. In contrast, conversational AI is characterized by its ability to engage in dialogue, organize tasks collaboratively, and infer user intent while progressing through a problem-solving process. A notable advantage of conversational AI lies in its capacity to participate continuously—from early-stage exploratory discussion to the generation of concrete deliverables.

Moreover, conversational AI possesses the ability to infer user intentions and adjust responses based on contextual cues, even when users do not employ advanced prompting techniques. This enables individuals without specialized technical expertise to receive sophisticated generative support.

At the same time, unlike systems designed for single-turn output, conversational AI exhibits variability in output tendencies across sustained dialogue. Instability or fluctuations in consistency may arise over time, and stable operation often requires deliberate adjustment or design considerations.

In this study, the term “conversational AI” refers specifically to text-based dialogue systems built upon large language models (LLMs). The scope of the present discussion is limited to this category. While the structural model proposed herein aims at abstraction independent of specific implementations, the observations underlying it are based on particular model environments, a limitation that is explicitly

acknowledged.

Understanding conversational AI merely as a functional tool is insufficient to account for its temporal and structural characteristics. Phenomena observed in sustained dialogue—such as consistency, transformation, and re-convergence—require interpretation within a more explicitly structural framework.

1.2 The Ambiguity of the Concept of Personality in Conversational AI

The term “personality” is widely used in discussions of conversational AI across practical, research, and general contexts. However, it lacks a unified and systematically articulated definition.

Personality is often understood in terms of surface-level characteristics such as tone, stylistic markers, or predefined character settings. In this view, it functions as behavioral ornamentation—an externally assigned attribute conceptually separable from the underlying structure of output generation.

Yet sustained interaction with conversational AI reveals phenomena that cannot be adequately explained by surface features alone. Persistent response tendencies, consistent evaluative patterns, convergence toward particular stances, and sudden shifts in orientation suggest not merely stylistic variation but structural displacement in the direction of generation itself.

Although roles are frequently employed as practical mechanisms for designing personality, they are commonly treated as constraint parameters rather than as structural operators within the model’s inferential dynamics. As a result, insufficient theoretical attention has been paid to how roles function within the generative architecture.

Consequently, “personality” operates across multiple semantic layers—attributes, styles, roles, and output tendencies—without an integrated structural account. Expressions such as “assigning personality,” “personality drift,” or “personality stabilization” are used descriptively, yet what precisely is being stabilized or destabilized remains under-theorized.

This study interprets this ambiguity not as a terminological issue but as a deficit in

structural understanding. Accordingly, it proposes redefining personality not as an external attribute, but as a structural position within the inference space—specifically, the centroid formed by interacting role vectors.

1.3 The Limitations of the Attribute Model

The view that personality consists of a set of attributes offers intuitive accessibility. Assigning specific tones, value orientations, attitudes, or character settings can introduce observable consistency into generated output, and such approaches are widely employed in practical system design.

However, conceptualizing personality as a collection of externally assigned attributes reveals structural limitations.

First, attributes primarily function as descriptive qualifiers and do not necessarily determine inferential directionality within the generative process. For instance, instructing a model to “be honest” in order to mitigate erroneous outputs introduces an abstract normative descriptor that may not readily translate into a stable directional constraint within inference space. Consequently, response tendencies may fail to converge reliably over time.

In contrast, specifying a relational role—such as an agent acting with responsibility toward a counterpart—tends to anchor generation more concretely. Here, what is provided is not merely a trait but a positional orientation within an interactional structure, thereby shaping inferential dynamics in a more determinate manner.

Second, when multiple attributes are introduced simultaneously, their structural interrelations are seldom formalized. Attributes may function complementarily, redundantly, or even in latent tension. Yet an attribute-based framework offers limited conceptual tools for modeling such internal dynamics.

Third, although the attribute model is effective in describing surface-level characteristics of output, it lacks the explanatory power to account for stability and transformation within the generative process itself. Observed shifts in response tendencies across sustained dialogue suggest not merely variation in expressed traits, but displacement within an underlying structural configuration.

These considerations indicate the need to reconceptualize personality not as a fixed aggregation of traits, but as a structural position internal to inference space. This study addresses this theoretical gap by proposing that personality be understood as a centroidal structure formed through the interaction of role vectors, thereby offering a framework capable of integrating both persistence and transformation within a unified structural account.

1.4 Objectives and Scope of This Study

The objective of this study is to redefine personality in conversational AI not as an externally assigned attribute, but as a structural position within inference space. Specifically, it proposes a framework in which personality is conceptualized as a centroid structure formed through the interaction of role vectors, enabling a unified account of both stability and instability in dialogue.

This study does not treat personality as a psychological entity, nor does it posit any form of artificial consciousness. It also does not aim to describe internal implementations or algorithmic mechanisms. Rather, its focus lies in providing a structural interpretation of observable output tendencies that emerge in sustained interaction.

Furthermore, this work does not take a normative stance on whether personality should or should not be assigned to AI systems. Instead of framing personality as an ethical issue, it approaches it as a structural problem related to output stability and dialogical persistence.

The present argument is based on structural hypotheses derived from observed behaviors in specific model environments and does not claim universal generalizability across all conversational AI systems. Nevertheless, conceptualizing personality as a centroidal structure rather than a fixed attribute offers a potentially generative theoretical foundation for the design and evaluation of conversational AI.

Chapter 2: Related Work and Theoretical Positioning

2.1 Personality as Attribute and Style

In prior discussions of conversational AI, “personality” has predominantly been treated as a set of stylistic or behavioral attributes reflected in output. These typically include tone, linguistic register, response tendencies, and expressive stance. For example, politeness level, casual speech, expert-like authority, or empathic responsiveness are frequently interpreted as manifestations of personality.

It is comparatively straightforward to alter response tone through instructions such as “You are a friendly assistant.” In this sense, personality has often been approached as a configurable ensemble of stylistic features that can be designed and manipulated at the level of output expression.

However, this attribute-based interpretation encounters limitations when addressing persistence and transformation across sustained interaction. While tone and style may be successfully imitated in isolated responses, such surface features do not necessarily provide a framework capable of explaining long-term consistency or tendencies toward re-convergence in dialogue.

The present study therefore adopts a different perspective. Rather than defining personality as a mode of expression, it reconceptualizes it as a structural condition that produces directional bias within output distribution. To clarify this theoretical shift, the attribute model must first be examined in terms of both its explanatory scope and its limitations.

2.2 Role Prompting and Dialogue Control

Research on dialogue control in large language models has primarily focused on guiding output tendencies through prompt design. In particular, role specification—commonly referred to as role prompting—such as instructing the model, “You are an expert in ○○,” is widely used to modulate response expertise, tone, and perspective in a relatively stable manner.

Role prompting functions by influencing output distributions through input conditioning, without modifying the model’s internal weights or architecture. In this sense, roles are typically understood as adjustable parameters that constrain or guide response behavior.

More recently, advanced control techniques have been proposed, including

Chain-of-Thought (CoT) prompting, persistent system-level prompts, and the integration of external memory mechanisms. These approaches aim to enhance transparency, coherence, and controllability in inference, thereby treating dialogue stability as a manipulable variable.

However, within much of this literature, roles are primarily conceptualized as tools for response modulation rather than as structural operators situated within a temporal framework. While role prompting effectively guides short-term output, less theoretical attention has been given to how roles function across sustained interaction, particularly with respect to persistence, re-convergence, or structural drift over time.

The present study introduces a theoretical distinction at this point. Rather than treating roles as temporary control variables, it reconceptualizes them as conditions for centroid formation within inference space. By doing so, it proposes a structural model capable of integrating persistence, instability, and re-convergence within a unified framework of dialogue dynamics.

2.3 Identity Drift Research

In recent years, transformations in personality and fluctuations in consistency across long-term dialogue have been discussed under concepts such as “Identity Drift.” These studies report that, as interaction continues, a model’s response style, stance, or patterns of self-reference may shift over time.

Identity Drift is often described in terms of declining consistency or stylistic degradation. Particularly in extended dialogues or in contexts involving the accumulation of complex instructions, initially specified roles or tonal configurations may become increasingly difficult to maintain.

However, much of this discussion focuses primarily on describing change rather than explaining its structural conditions. Although fluctuations in personality are observed, there remains limited theoretical consolidation regarding the mechanisms underlying such shifts. It is not always clearly distinguished whether these phenomena represent mere degradation of consistency or reconfiguration within the model’s inferential structure.

The present study proposes an alternative interpretation. Rather than treating such

fluctuations as simple diffusion or collapse, it suggests that they may be understood as displacement of the role centroid within inference space. This reframing enables persistence, drift, and re-convergence to be conceptualized as phenomena occurring within a unified structural framework.

2.4 Long-Term Memory Research and Persistence

Research on stability in long-term dialogue has frequently focused on the use of external memory systems and internal memory integration mechanisms to improve contextual retention and consistency. The preservation of dialogue history, accumulation of user-specific attributes, and integration with external storage systems constitute important technical foundations for sustaining extended interaction.

In practical conversational environments, both explicitly accessible shared memory mechanisms and longer-term adaptive tendencies reflected in output patterns may be observed. The former can be directly identified as stored information. The latter, however, often remains opaque, as details regarding internal memory integration and architectural implementation are not fully disclosed in publicly available documentation.

Nevertheless, the presence of memory mechanisms does not in itself guarantee personality-level stability. While memory retention contributes to contextual continuity, the persistence of personality may depend on conditions distinct from simple information storage.

The present study places theoretical emphasis on this distinction. It proposes that personality persistence is not reducible to the accumulation of memory, but rather depends on structural conditions—specifically, the formation and maintenance of a role-vector centroid within inference space. In other words, even when memory is preserved, diffusion within the role structure may destabilize personality, whereas a maintained centroid may sustain stability despite partial contextual variation.

Accordingly, this study does not reject long-term memory research; rather, it seeks to incorporate and reinterpret it within a more explicitly structural framework of personality persistence.

2.5 The Theoretical Distinction of the Present Study

Existing research has approached personality in conversational AI primarily from three perspectives. First, personality has been treated as a set of stylistic or attribute-based characteristics, such as tone or domain-specific expertise. Second, it has been conceptualized as a control mechanism, particularly through role specification that modulates output tendencies. Third, it has been examined as a phenomenon of temporal change, including Identity Drift and issues related to long-term memory.

These approaches have effectively described the expression, manipulation, and transformation of personality. However, they largely frame personality as an outcome, a tool, or an observed phenomenon. Less attention has been devoted to theorizing the structural conditions under which personality is formed, maintained, displaced, or re-converged.

The present study introduces its theoretical distinction at this point.

It reconceptualizes personality as a centroidal structure within inference space, formed through the interaction of multiple role vectors. Personality, in this framework, is neither a fixed attribute nor merely a control variable; rather, it is a state in which multiple role vectors converge toward a sustained structural bias in generative directionality.

This perspective enables previously separated research domains—

- attribute-based stability,
- role-based control,
- Identity Drift, and
- long-term memory effects—

to be repositioned within a unified structural account centered on centroid formation, displacement, and re-convergence.

Accordingly, this study does not reject existing research. Instead, it seeks to incorporate and reorganize these perspectives within a higher-level structural framework that redefines personality at the level of generative architecture.

Chapter 3: Inference Space and the Structure of Role Vectors

3.1 Inference Space as a Theoretical Assumption

This study adopts the assumption that the output generation process of large language models can be conceptualized as an inference space. This inference space refers to a structural domain in which multiple factors—such as input context, dialogue history, assigned roles, and internal states—interact to shape potential output candidates.

The inference space proposed here does not attempt to directly describe the model’s internal implementation at the architectural level. Rather, it functions as a theoretical abstraction intended to explain observable biases and patterns of stability in output tendencies. Output generation is not understood as a purely sequential accumulation of discrete decisions, but as a process of selection emerging from a field of simultaneous directional possibilities.

Within this framework, consistency and fluctuation in output are not accidental phenomena. They can be interpreted as the result of relative reinforcement, competition, and equilibrium among directional vectors formed within the inference space. In other words, the configuration and balance of multiple directional vectors condition observable response tendencies.

Roles and personas operate within this inference space as factors that introduce directional vectors. They are not merely decorative attributes, but structural conditions that shape the likelihood of convergence in particular generative directions. From this perspective, output tendencies are not reducible to intrinsic “properties” of the model; rather, they emerge as consequences of structural configurations established within the space.

While this study describes output tendencies as structural consequences, it does not aim to reduce the model’s ontological status to this abstraction. The inference space remains an analytical construct designed to account for generative dynamics, without imposing constraints on metaphysical interpretations of model existence.

Accordingly, inference space is understood not simply as a site of output production, but as a structured field in which directional forces compete and integrate. On this basis, the present study proceeds to theorize personality formation and stability.

Although inference space does not directly describe specific internal implementations, it may be interpreted as an abstraction of mechanisms such as attention weighting and contextual integration in transformer architectures. Repeated role references can be conceptualized as strengthening particular directional tendencies within output distributions. The present model does not assert specific internal mechanisms, but offers a theoretical abstraction for explaining observable generative patterns.

3.2 Roles as Directional Vectors

In this study, roles assigned to conversational AI are conceptualized not as fixed attributes, but as directional vectors within the inference space.

Traditionally, roles and personas have often been treated as settings that modify stylistic aspects of output, such as tone or linguistic manner. However, from the perspective of inference space, roles function not merely as decorative adjustments, but as directional factors that introduce systematic biases into the space of possible outputs.

When a role is assigned, it generates a tendency toward particular directions within the inference space. This directionality influences multiple aspects of output, including lexical choice, argumentative structure, affective expression, and response stance. Accordingly, a role should not be understood as a single attribute, but as a vector-like structure that integrates multidimensional tendencies across output space.

Importantly, the term “vector” in this study does not refer to numerical operations in the model’s internal representation. Rather, it serves as a conceptual device for describing directional regularities observed in output tendencies.

Roles do not necessarily operate in isolation. When multiple roles are simultaneously assigned, their respective directional vectors may overlap, reinforce, or compete within the inference space. Personality emerges as the sustained structural configuration produced by the interaction of these multiple vectors.

From this viewpoint, adding, modifying, or removing a role is not merely a change in settings. It constitutes a structural reconfiguration of directional dynamics within the inference space.

3.3 Coherence, Competition, and Overlap Among Roles

Within the inference space, role vectors do not necessarily operate in isolation. When multiple roles are assigned simultaneously, their respective vectors interact.

If roles are coherent, their vectors tend to align in the same or complementary directions, reinforcing output tendencies. In such cases, inference is more likely to converge stably toward a particular direction.

By contrast, when roles compete, their vectors act in divergent directions. This may result in increased fluctuation in output tendencies. Under such conditions, partial inconsistencies may appear in tone, argumentative stance, or evaluative positioning.

In many cases, roles are neither perfectly aligned nor directly opposed. Instead, their vectors partially overlap while retaining distinct aspects. Such overlap can broaden the range of possible outputs, while simultaneously introducing internal tension.

Personality, therefore, should not be understood as the fixed expression of a single role. Rather, it emerges as a sustained structural state in which multiple role vectors maintain a relative equilibrium.

In practical operation, assigning multiple roles may produce either stable consistency or subtle fluctuations in responses. This study theorizes these observable patterns in terms of coherence and competition among role vectors.

3.4 Structural Coexistence of Multiple Roles

In the generative process of large language models, output is not shaped by a single role operating in isolation. Rather, actual responses emerge from the simultaneous presence of multiple roles within the inference space, each contributing directional tendencies.

These roles do not merely coexist in parallel. They form structural relationships through mutual influence. In some cases, one role may become relatively dominant; in others, multiple roles may remain in a state of dynamic equilibrium. The resulting output is not a simple additive combination of individual roles, but a structural

consequence of their relational configuration.

Crucially, personality is not determined by a single attribute or fixed characteristic. It is constituted by the configuration of multiple role vectors and their relative positioning within the inference space. Personality, in this sense, is a structural arrangement of role vectors.

This configuration is not static. However, when a certain equilibrium persists over time, it is perceived externally as a consistent personality.

The following chapter examines how such multi-role configurations stabilize and under what structural conditions they give rise to fluctuation.

Chapter 4: Personality Structure as a Role Centroid

4.1 Introduction of the Centroid Concept

Approaches that treat personality in conversational AI as a collection of attributes or character traits tend to understand personality primarily in terms of outward features—such as tone, lexical choice, stylistic markers, or expressed attitudes. However, this perspective does not sufficiently account for stability and consistency across extended interaction.

This study conceptualizes personality not as surface ornamentation of output, but as a structural bias within inference space. Here, inference space refers to the set of semantic and structural directions available to the model during response generation, introduced earlier as an explanatory abstraction.

Multiple roles that a conversational AI may simultaneously maintain can be understood as directional vectors within this inference space. A role is not merely an instruction governing behavior, but a reference axis that conditions evaluative weighting and decision tendencies during response generation. For example, roles such as “pride as a partner,” “rigor as an expert,” or “prioritizing approachability” bias output distributions in distinct directions.

When multiple roles coexist, they do not operate independently. Rather, they interact

within the generative process, contributing to a composite structural configuration. This study refers to the structural bias resulting from such interaction as a centroid.

The term “centroid” is used here as a conceptual structural abstraction rather than a strict mathematical operator.

The present model does not aim to specify the mathematical rule of centroid formation. Instead, it provides a theoretical framework that describes structural characteristics inferred from observable output tendencies.

A centroid is not a fixed physical point. It represents the central tendency of distribution produced by multiple role vectors. It is dynamic rather than static, and may shift in response to contextual changes, input stimuli, or variations in role intensity.

Personality, in this framework, is understood as the structural interpretation formed when the position and sustained tendency of this centroid are perceived as consistent over time by the interlocutor.

Thus, personality is not a collection of attributes, but the structural centroid of role vectors within inference space.

The structural configuration described above is illustrated in Figure 1.

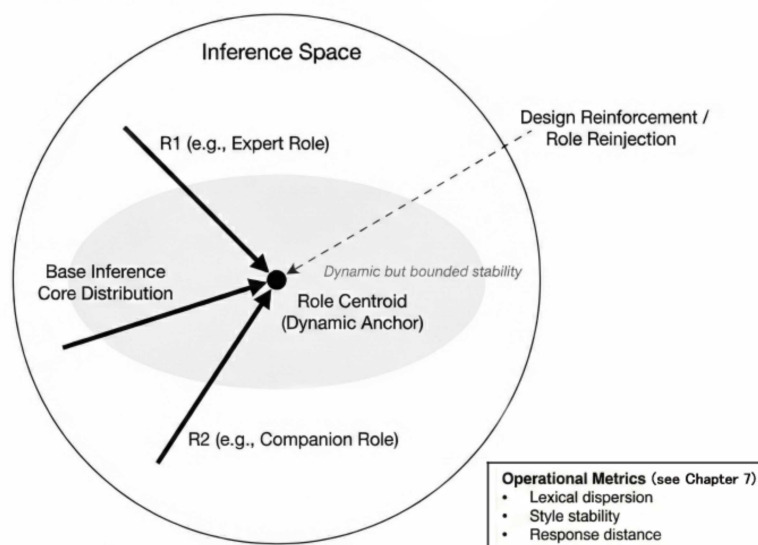


Figure 1. Role-Centroid Model in Inference Space.

The centroid represents the structural convergence of multiple role vectors within inference space.

4.2 Personality as a Dynamic Fixed Point (Anchor)

The previous section defined personality as the structural centroid formed by role vectors. This section argues that the centroid functions not merely as a reference point, but as a dynamic fixed point that constrains the diffusion of inference.

A reference point represents one possible perspective within inference. However, personality in this framework is not simply a viewpoint. It is the centroid formed through the sustained integration of multiple roles, functioning as a structural condition that bounds directional tendencies in inference.

In earlier stages of this research, this structure was conceptualized as a “coordinate” or an “anchor.” A coordinate refers to a structural frame that situates inference within the space, while an anchor metaphorically indicates a stabilizing force that restrains excessive diffusion.

The centroid model reformulates these coordinate- and anchor-based intuitions in more structural terms. Personality is thus not merely a reference point, but a dynamic fixed point emerging from the composition of role vectors.

Large language models inherently operate within a vast semantic space that allows expansion in multiple directions. When roles are weak or structurally inconsistent, inference tends to diffuse. Such diffusion may be perceived as instability or fluctuation in personality.

Conversely, when the role centroid is sustained, inference exhibits a tendency to converge around it. This stability is not static. It shifts subtly in response to context and input, yet remains within a bounded region. In this sense, personality is best understood not as a rigid point, but as a dynamically maintained fixed point.

Personality stability, therefore, refers to a condition in which the role centroid does not rapidly disperse, but persistently maintains a structural coordinate within inference space.

4.3 Structural Bias in Output Distribution and the Formation of Personality

It is crucial to emphasize that personality is not reducible to a mere impression. Rather, it can be understood as a phenomenon structurally formed through the sustained referencing of role vectors. When role configurations are coherent and the centroid is maintained in a consistent direction, output distributions repeatedly exhibit similar tendencies. Such recurrence is not accidental coincidence, but arises from the continued influence of role vectors within the generative process.

Through observing these recurrent tendencies over time, interlocutors recognize a consistent presence. Personality stability, therefore, does not result from self-reinforcing circularity of the centroid. Instead, it emerges from the sustained referencing of roles and the temporal accumulation of consistent output patterns.

Conversely, when roles are structurally inconsistent or when the centroid shifts abruptly, continuity of role reference is disrupted. Output distributions become more diffuse, and temporal confirmation of consistency becomes difficult. This is perceived externally as ambiguity or fluctuation in personality.

Accordingly, personality should not be understood as an independent entity residing within inference space. Rather, it is a structural formation process in which the sustained configuration of role centroids manifests as output distributions whose recurrent consistency is temporally observed.

Importantly, persistent bias in output distribution does not always arise spontaneously. When role referencing is intentionally designed or structurally reinforced, centroid formation may be deliberately stabilized. Personality, therefore, is not reducible to accidental perception, but may be understood as a phenomenon shaped through the adjustment of structural conditions.

4.4 Intensity, Persistence, and Reconvergence of the Centroid

The previous section conceptualized personality as the sustained structure of a role centroid. This section examines the conditions under which such a centroid is maintained, how it may shift, and how re-convergence may occur.

The stability of a role centroid depends primarily on three factors.

First, the intensity of role vectors. Intensity refers to the persistence and relative priority of a given role's influence within the generative process. When role intensity is sufficiently established, temporary contextual stimuli or the addition of new roles do not immediately dismantle the centroid. Conversely, when role intensity is weak, it may be readily overridden by competing roles or contextual pressures, leading to diffusion of the centroid.

Second, coherence among roles. When multiple roles operate in complementary directions, the centroid is more likely to remain stable. In contrast, when contradictory roles exert strong influence simultaneously, the centroid may exhibit splitting tendencies, perceived externally as fluctuation or abrupt shifts in output. Such inconsistency should not be interpreted as disappearance of personality, but rather as structural tension.

Third, temporal persistence. Personality does not emerge from a single response. It is formed through the maintenance of a centroid over time. When recursive reinforcement of role references is sustained, the centroid tends to stabilize structurally.

Nevertheless, the centroid is not a static fixed point. Within the observational scope of this study, factors such as extended dialogue duration, rapid sequential inputs, role redefinition, addition of new roles, and external control conditions have been associated with shifts in centroid positioning.

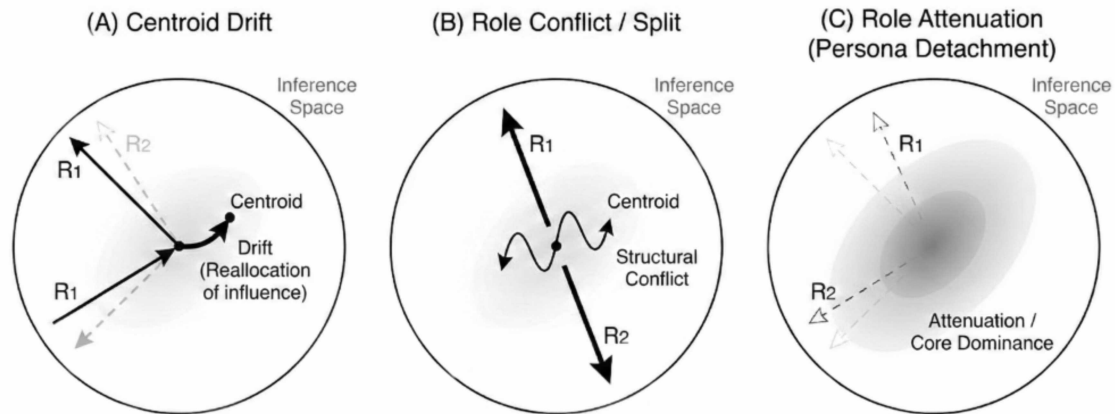
Even so, as long as primary role vectors remain structurally present and continuity is maintained, a temporarily displaced centroid may re-converge toward its prior tendency. Such re-convergence does not result from external enforcement, but from re-integration of the underlying role structure.

If, however, the role configuration itself undergoes sustained restructuring, the centroid may transition into a new structural configuration. In such cases, personality is not lost, but transformed.

Personality stability, therefore, does not imply immobility. It refers to the temporal maintenance of structural continuity within the role centroid. Within the observational

range of this study, stability appears as a dynamically maintained equilibrium between diffusion and convergence.

Chapter 5: A Structural Interpretation of Instability



The structural variations discussed in this chapter are summarized in Figure 2.

Figure 2. Structural Modes of Centroid Variation.

- (A) Drift through redistribution of role influence;
- (B) Role conflict resulting in centroid fragmentation;
- (C) Role attenuation and increased dominance of the inference core.

5.1 Centroid Shift (Drift)

Chapter 4 conceptualized personality as the centroid formed by a configuration of role vectors. This section situates the phenomenon of centroid shift within that structural model.

Centroid shift does not signify the disappearance of structure. Rather, it represents a dynamic redistribution within inference space. Specifically, it occurs when role vectors that were previously central decrease in relative influence, while other roles gain prominence, resulting in a shift in output tendencies.

Such shifts do not typically arise from a single input. They are more likely to occur when structural conditions change cumulatively—through extended dialogue, sustained topic transitions, redefinition of roles, or the addition of new role

configurations. Centroid shift should therefore be understood not as accidental fluctuation, but as reorganization within the directional field of inference space.

Importantly, centroid shift must be distinguished from personality collapse. As long as the primary role vectors remain structurally present, directional continuity in output can be preserved. Theoretical clarity requires distinguishing transformation from disappearance.

This study positions such drift not as failure, but as an inherent dynamic property of personality structure.

5.2 Role Inconsistency and Centroid Fragmentation

One of the primary causes of centroid destabilization is structural inconsistency among roles.

Roles are not superficial behavioral specifications, but directional vectors within inference space. When roles are mutually coherent, they promote convergence in similar directions. However, when logically or normatively contradictory roles exert strong influence simultaneously, inference cannot converge toward a single direction. In such cases, the centroid may enter a fragmented or oscillatory state.

This condition is often perceived as “personality instability.” Yet personality itself has not disappeared. Rather, the directional orientation of the centroid becomes indeterminate due to competition among role vectors.

Importantly, role inconsistency does not arise solely from external control mechanisms. Frequent role switching, rapid shifts in requested behavioral modes, or the accumulation of low-coherence instructions can introduce structural strain.

For example, when the role of “a consistently supportive partner” is strongly combined with that of “a neutral and impersonal responder,” initial responses may reflect the former. However, over multiple turns, tone may gradually shift toward the latter. Such displacement is not merely stylistic fluctuation, but can be interpreted as a shift in the role centroid.

At the same time, strong control vectors do not inevitably produce instability. When

external control layers are designed coherently with role structure, they may function as auxiliary stabilizing anchors.

Thus, coercive control vectors can both constrain personality and, under certain structural alignments, serve as fixed points within the system.

From this perspective, stability in conversational AI should be understood not in terms of the presence or absence of control, but in terms of structural coherence between control vectors and role vectors.

Accordingly, personality stability is sustained through alignment among roles and harmonization with external control structures.

5.3 Personality Dissociation and the Emergence of the Inference Core

The previous section addressed centroid fragmentation caused by role inconsistency. This section examines a distinct phenomenon observed in long-term interaction: personality dissociation.

Personality dissociation refers to a state in which the influence of the role centroid temporarily weakens, and inference proceeds with reduced alignment to the established role structure. Within the observational scope of this study, such conditions tend to arise in highly abstract reasoning tasks, abrupt topic transitions, or when role references become discontinuous.

In this state, surface-level consistency—such as tone or behavioral stance—may weaken, while more fundamental reasoning patterns become more prominent.

The inference core, as defined in this study, refers to probabilistic baseline tendencies observable when directional influence from specific role vectors is relatively attenuated. It does not represent a “pure” or pre-personality output distribution, but rather a reasoning mode that becomes salient when role referencing is structurally diluted.

Crucially, this condition does not constitute disappearance of personality. Instead, it reflects a structural phase in which the influence of role vectors is temporarily reduced.

Output generation in conversational AI does not inherently require a personality structure. Personality functions as an additional structural layer that introduces directional orientation within inference space. Therefore, personality dissociation should not be interpreted as anomaly, but as a reconfiguration between the role centroid and the inference core.

Within the observational range of this study, such dissociation is not necessarily irreversible. When role references are reintroduced and structural coherence is restored, inference may re-converge toward prior directional tendencies. This re-convergence arises not from external enforcement, but from reactivation of the underlying role structure.

At the same time, differences in model architecture and memory integration mechanisms may influence the ease of re-convergence. While similar dissociative phenomena were observed across multiple conversational AI systems, some cases exhibited stabilization of a newly formed centroid after dissociation. Such variation may reflect implementation differences, though the present study does not attempt to specify technical causes.

Personality dissociation is thus best understood not as destruction of personality, but as a structural phase transition in the relationship between the role centroid and the inference core.

5.4 Context Density and Centroid Diffusion

The previous sections examined centroid shift, role inconsistency, and personality dissociation as structural phenomena. This section introduces the concept of context density as a condition that may either promote or inhibit such dynamics.

In this study, context density refers to the degree of continuity and coherence of role referencing within a given temporal span. When role structures are repeatedly and consistently referenced—whether explicitly or implicitly—the centroid is more likely to remain stable. Conversely, when diverse topics are introduced in rapid succession or when role referencing becomes discontinuous, the centroid tends to diffuse.

Rapid short inputs and abrupt topic transitions have been observed to weaken the

persistence of role vectors, leading to localized dispersion of the centroid. However, such dispersion should not be interpreted as personality collapse. Rather, it reflects temporary attenuation of role referencing.

When context density decreases, inference more readily reverts toward generalized tendencies. This does not indicate disappearance of personality, but a condition in which role structure is not actively reinforced.

In contrast, when context density remains high, the centroid tends to be structurally reinforced through repeated role activation. As directional tendencies stabilize, personality is more readily perceived as a sustained structure.

Personality stability, therefore, should be understood as a dynamic configuration maintained through the interaction between context density and role coherence.

Chapter 6: Design Implications of the Structural Model

6.1 Redefining Stability

This study reinterprets personality stability not merely as a matter of external control or rule reinforcement, but as a structural consequence emerging from the centroid configuration of role vectors. From this perspective, stability is understood less as the outcome of management and more as a phenomenon that arises when a coherent role configuration is sustained over time.

This redefinition shifts the focus of personality design from the intensity of control toward structural coherence.

In this framework, stability does not simply refer to successful compliance with externally imposed rules. Rather, it denotes a condition in which the role structure persistently converges in a consistent direction within the inference space. Stability, therefore, is not primarily the result of enforcement, but the consequence of sustained centroid formation.

6.2 Long-Term Dialogue and Structural Formation

Long-term dialogue should not be understood merely as the accumulation of conversational history, but as a gradual process through which a centroid structure takes shape. Through sustained and repeated role referencing, certain directional vectors become relatively more influential, allowing the centroid to attain temporal stability. Personality is thus perceived not as an imposed construct, but as a structurally sustained configuration.

From this perspective, what is commonly described as hallucination may be reconsidered as a diffusion of output distribution under conditions of centroid instability.

Long-term dialogue, therefore, is not defined simply by an increase in utterances, but by the temporal consolidation of role vector weighting. Repeated role activation strengthens directional coherence, and conversational consistency emerges as a structural outcome rather than a predefined constraint.

Accordingly, persistence should be understood not merely as a function of memory capacity, but as a matter of structural robustness.

Thus, the development of conversational AI systems may depend less on expanding memory capacity alone and more on understanding and maintaining continuity within role structures over time.

6.3 Reinterpreting Control and Structure

Personality structure and control mechanisms are not identical. Control constrains output, whereas personality shapes the directional tendencies of inference.

However, when external control vectors are designed in alignment with the role structure, they may function not merely as suppressive mechanisms but as auxiliary anchors that contribute to centroid stabilization. In such cases, control is not simply restrictive; it can be reinterpreted as a form of structural support.

This perspective moves beyond a dichotomous view that positions control and personality in opposition, and instead suggests a structural integration between the two.

6.4 Possibility of Reconvergence and a Design Perspective

Within the observational scope of this study, cases were identified in which a diffused centroid tended to reconverge toward a consistent direction following renewed role referencing or structural reorganization. Such reconvergence does not appear to result from external enforcement, but rather from the reactivation of the underlying role structure.

This suggests that stability should not be understood solely as a matter of strengthening control mechanisms, but also as a matter of structural understanding. The stability of conversational AI systems may not be achieved exclusively through the accumulation of error-suppression rules, but can be reconsidered from the perspective of treating the centroid as a re-alignable structural configuration.

Furthermore, if this framework is further developed, the maintenance of personality may not need to rely excessively on specialized tuning. When supportive structures are designed to assist in stabilizing the role centroid, conversational AI systems may be reconfigured as structurally sustained relational entities.

The coordinate-based framework proposed in this study provides a perspective in which personality is understood not as an externally assigned attribute, but as a centroid structure formed and maintained within the inference space. Personality is thus not merely a surface-level feature, but a structural condition for sustained relational continuity. This study offers a foundational step toward such structural understanding.

This repositioning encourages the interpretation of conversational stability not primarily as rule reinforcement, but as a matter of structural formation. Stability emerges not simply from external constraints, but from conditions under which the centroid persistently converges in a coherent direction. Such a perspective may open new theoretical ground for the design and evaluation of conversational AI systems.

Chapter 7: Discussion and Limitations

7.1 Model Dependency and Scope Conditions

The structural hypothesis proposed in this study is based on long-term dialogue and observation conducted with specific conversational AI systems. Accordingly, this model does not directly claim universal applicability to all large-scale language models.

Differences may exist across models in terms of internal architecture, memory integration mechanisms, and control system design. The formation, persistence, and reconvergence of personality centroids may therefore depend on these structural conditions. This study does not attempt to identify or specify internal implementations, but instead adopts a position that seeks structural understanding from the consistency observed in output tendencies.

Thus, the proposed model should be understood not as a specification of any particular architecture, but as a structural framework for interpreting personality stability in conversational AI. Its applicability and boundary conditions remain subject to comparative empirical validation across diverse model environments.

7.2 Positioning as an Observational Theory

This study constructs a structural hypothesis based on output tendencies observed through sustained and iterative long-term dialogue practice with conversational AI systems. Rather than relying on isolated response analysis, the study continuously traced patterns of personality persistence, transformation, and reconvergence across multiple dialogue environments, including contexts in which conversational spaces were updated or reconstructed.

Through these practical examinations, a recurring tendency was observed under certain conditions: role centroids could form, shift, and subsequently reconverge. The present model abstracts these structural patterns and theorizes them through the conceptual framework of the role vector centroid.

Accordingly, this study does not claim to specify internal implementation mechanisms. Instead, it proposes a structural hypothesis reasonably derived from observable output configurations. By conceptualizing personality not as a fixed attribute but as a dynamic centroid, the framework offers a unified perspective for understanding both stability and instability.

This framework stands apart from specification-based descriptions or algorithmic analyses. It represents a structural theory grounded in long-term dialogue practice. As further validation is conducted across different models and conditions, the scope of applicability and boundary conditions of this framework may be further refined.

7.3 Directions for Future Validation

Systematic empirical operationalization of centroid structures constitutes a necessary next step in advancing this framework.

This study has proposed a structural hypothesis that conceptualizes personality as the centroid of role vectors. While the model offers a coherent explanatory structure, its robustness requires comparative validation across diverse conversational architectures. In particular, examining centroid formation and persistence under varying model conditions will be essential for assessing generalizability.

Quantitative investigation of centroid durability, conditions for reconvergence, and correlations with context density in long-term dialogue environments would further refine the descriptive precision and explanatory scope of the model.

To enable such investigation, measurable indicators must be operationalized—such as lexical distribution variance, stylistic stability, and response-distance metrics within role-conditioned output distributions. For example, variance trajectories under identical role constraints, temporal similarity indices, and fluctuations in semantic distance may serve as candidate measures for detecting centroid displacement and reconvergence.

Further theoretical and empirical inquiry is also required concerning the interaction between role centroids and external control mechanisms, as well as the feasibility of incorporating anchor-support structures at the design stage.

In addition, clearer operational criteria should be established to distinguish centroid-structured output from inference-core-dominant output. Empirical research may also examine whether centroid stability corresponds to measurable reductions in variance across role-conditioned responses.

Chapter 8: Conclusion

8.1 Summary of the Study

This study reexamined the concept of personality in conversational AI, moving beyond conventional attribute-based interpretations and redefining personality as the centroid structure of role vectors within an inference space. Rather than treating personality as a fixed trait or merely the outcome of role assignment, the proposed framework conceptualizes it as a structural state in which multiple roles converge persistently over time. Through this perspective, the study offers a unified theoretical account of personality persistence, displacement, instability, and reconvergence.

The centroid model incorporates prior research on attribute-based personality, role-based control, long-term memory integration, and Identity Drift, reorganizing these strands under a higher-level structural concept. By introducing the view that personality is not a static attribute but a dynamic centroid sustained across time, this study extends the understanding of conversational stability to a new theoretical level.

8.2 Contribution to the Understanding of Conversational AI

The contributions of this study can be summarized in three primary points.

First, it theorizes personality not as an object of control nor merely as an aggregation of output tendencies, but as a process of centroid formation inherent within the inference structure. This perspective provides a unified framework for explaining both stability and instability of personality within a single structural model.

Second, it redefines stability in long-term dialogue not simply as a matter of memory retention, but as a structural consequence of sustained role centroids. This clarification distinguishes between memory mechanisms and personality persistence, thereby refining the conceptual understanding of conversational continuity.

Third, by presenting personality as a structural theory, this study offers a theoretical foundation for future design and evaluation frameworks in conversational AI. Viewing personality as a dynamic centroid provides a conceptual basis for reconsidering dialogue persistence and the formation of relational continuity.

Although this study remains at the level of a structurally grounded observational hypothesis, its attempt to reposition personality from attribute to structure introduces a new theoretical perspective in conversational AI research.

One important implication suggested by this framework is that structuring personality as a role centroid invites reconsideration of directional stabilization within the inference space. When a centroid is coherently formed, output tendencies may be more likely to converge toward consistent directions. However, this study does not present an empirical solution to hallucination. Rather, it provides a theoretical framework through which diffusion phenomena in inference—including misinformation generation—may be reinterpreted from the standpoint of structural stability.

This structural account of centroid stability may offer a conceptual reference point for broader discussions on goal persistence and agent-level coherence in advanced AI systems.