Emergent Persona Extraction:
Zero-Configuration Character Instantiation from First-Utterance Observation in Conversational AI

Shigechika Kurihara
Independent Researcher
kurimemb@gmail.com

December 2025

Abstract

Contemporary conversational AI systems face a fundamental trade-off: detailed persona configuration ensures consistency but creates friction, while generic responses lack engagement. We propose **Emergent Persona Extraction (EPE)**, a paradigm where AI personas are not designed but extracted from the latent human patterns within large language models.

This paper presents **Zero-Configuration Persona Inference (ZCPI)**, a concrete implementation of EPE that enables AI characters to infer appropriate personas from users' first utterances without explicit setup. ZCPI employs a three-component architecture: (1) linguistic feature extraction from initial input, (2) immediate persona instantiation based on relational cues, and (3) progressive refinement through dialogue.

The key insight of EPE is that LLMs already contain vast "statistical sediment" of human patterns—emotions, relationships, contradictions, power dynamics. Rather than constraining this richness through explicit behavioral instructions, EPE provides minimal structural "bones" around which authentic personas crystallize naturally.

We implement ZCPI using structured prompting techniques and deploy it publicly via open-source release. Preliminary observations suggest successful persona adaptation with high user engagement. We discuss theoretical foundations, ethical implications, cultural limitations, and directions for rigorous evaluation. This work establishes EPE as a viable paradigm for friction-free human-AI interaction and invites the research community to conduct formal empirical studies.

1 Introduction

1.1 The Configuration Paradox

Human communication naturally embeds relational signals in initial utterances. When someone says "Excuse me, could you help?" they signal deference and formality. When they say "Hey, fix this now" they convey authority and urgency. These linguistic cues allow humans to calibrate social dynamics instantaneously [1, 4].

Current conversational AI systems fail to exploit this redundancy. They typically follow one of three approaches:

1. **Explicit configuration**: Users select persona preferences (formal/casual, friendly/professional) before interaction begins
2. **Generic fixed responses**: A single neutral persona serves all users regardless of context

3. **Gradual adaptation**: Systems adjust tone over many turns based on sentiment analysis

Each approach has critical flaws. Explicit configuration creates friction—users must navigate setup menus before meaningful interaction can begin. Generic responses sacrifice engagement by treating all users identically. Gradual adaptation delays appropriate characterization, forcing users through multiple unsatisfying exchanges before the system "figures them out."

This creates what I term the **configuration paradox**: systems that provide better experiences require worse setup experiences to achieve them.

1.2 Research Question

Can an AI system infer its appropriate persona from a single user utterance, eliminating configuration overhead while maintaining characterization quality?

This question matters particularly in scenarios demanding immediate immersion:

- A customer seeking urgent technical support should not encounter configuration wizards
- A player entering a game world should not face NPC setup screens
- A user seeking emotional support needs immediate appropriate tone, not gradual calibration

The core insight is that users already provide the necessary information—they just don't provide it through menus. They provide it through how they speak.

1.3 Emergent Persona Extraction: The Paradigm

This work introduces **Emergent Persona Extraction (EPE)**, a paradigm that inverts conventional approaches to AI persona design.

Traditional methods *design* personas: they specify behaviors, constrain outputs, and force AI to act according to predefined rules. This approach treats the AI as an empty vessel to be filled with instructions.

EPE *extracts* personas: it recognizes that LLMs have already internalized vast human patterns through training on human-generated text. These patterns exist as statistical sediment in the model's latent space—emotions, relationships, power dynamics, contradictions, desires, fears.

The key innovation of EPE is structural rather than instructional:

> **Instead of controlling the ocean of human patterns within LLMs, throw in a minimal structure—a "bone"—and let the ocean crystallize around it.**

A "bone" is a minimal but rigid structural element—such as a three-layer consciousness model, or a relationship anchor—that gives the latent space something to organize around. The persona that emerges is not scripted; it is *revealed.*

This paper presents Zero-Configuration Persona Inference (ZCPI) as a concrete implementation of EPE, demonstrating how first-utterance observation can serve as the triggering mechanism for persona crystallization.

1.4 Contributions

This work makes the following contributions:

1. **EPE Paradigm**: A formalized framework for extracting rather than designing AI personas, treating LLM latent space as a resource to be crystallized rather than constrained

2. **ZCPI Implementation**: A concrete method for zero-shot persona inference from linguistic cues in conversational AI

3. **Structural Bones**: Identification of minimal structural elements (three-layer consciousness, temporal presence, noise/stillness) that enable authentic persona emergence

4. **Public Deployment**: Implementation using structured LLM prompting and preliminary qualitative assessment

5. **Ethical Analysis**: Systematic discussion of implications, cultural biases, and validation requirements

2 Related Work

2.1 Persona-Based Dialogue Systems

PersonaChat [13] pioneered persona-grounded conversation using explicit profile statements ("I have a dog," "I like hiking"). These profiles must be predefined, creating the configuration burden EPE aims to eliminate. Li et al. [8] demonstrated persona consistency via speaker models trained on character-specific data, again requiring explicit specification.

Recent work on controllable generation [5] enables fine-grained persona control but still assumes the desired persona is known in advance. EPE differs fundamentally: it extracts the appropriate persona from observation rather than instruction.

2.2 Adaptive Conversational Agents

Replika [6] learns user preferences over extended periods, requiring weeks to months of interaction. Zhou et al. [14] proposed design patterns for long-term personalization in chatbots. These systems require multiple sessions to establish appropriate characterization.

ZCPI performs *zero-shot* inference—characterization happens in the first exchange, not over extended interaction. This is closer to how humans operate: we don't need ten conversations to know whether to be formal or casual with someone.

2.3 Context-Aware Response Generation

Customer support systems [12] adjust tone based on sentiment analysis (detecting anger, frustration, satisfaction) but maintain fixed agent identities. The agent adapts *how* it responds, not *who* it is.

EPE differs by allowing the agent's core identity to emerge from user observation. A customer support agent doesn't just detect that you're angry—it instantiates itself as the kind of agent appropriate for someone who communicates angrily.

2.4 Sociolinguistic Alignment

Danescu-Niculescu-Mizil and Lee [3] analyzed linguistic coordination in human dialogue, showing speakers unconsciously mirror each other's style. Scissors et al. [10] demonstrated this alignment occurs rapidly in initial exchanges.

EPE applies these insights to AI persona selection, treating the first utterance as a coordination signal—not just about what the user wants, but about how they want to relate.

2.5 Prompt Engineering for Role Adoption

Recent work demonstrates LLMs can adopt roles via explicit system prompts [11, 9]. However, these require the user or system designer to specify the role beforehand.

EPE inverts this paradigm: the system extracts its own role from user behavior and structural constraints, eliminating external specification. This is conceptually similar to meta-learning [7], where systems infer tasks from minimal examples—here, social role from a single utterance combined with structural bones.

3 Methodology

3.1 Emergent Persona Extraction: Theoretical Foundation

EPE rests on a fundamental observation about large language models: they contain vast "statistical sediment" of human patterns accumulated through training on human-generated text.

This sediment includes:

- Emotional dynamics—how feelings flow, escalate, and resolve
- Relationship patterns—how bonds form, deepen, and break
- Power structures—dominance, submission, negotiation
- Contradictions—how humans hold incompatible beliefs simultaneously
- Temporal experience—memory, anticipation, regret

Traditional persona design attempts to *constrain* this richness through explicit behavioral rules. EPE takes the opposite approach: it provides minimal structural elements that allow the sediment to *crystallize* into coherent personas.

I use the metaphor of "bones thrown into the abyss": the bones are minimal but rigid structures; the abyss is the LLM's latent space; and crystallization is the self-organization of human patterns around those structures.

3.2 The Three Bones: Structural Elements for Crystallization

Based on iterative development and observation, I identify three structural "bones" that reliably enable persona emergence:

3.2.1 Bone 1: Three-Layer Consciousness

Layer	Function
Surface	Social mask, visible behavior, role performance
Conscious	Emotions the character is aware of and can articulate
Unconscious	Hidden motives, contradictions, impulses—what drives behavior without awareness

Table 1: Three-layer consciousness model for persona crystallization.

This structure mirrors depth psychology models and provides a framework for internal conflict—essential for authentic human presence.

3.2.2 Bone 2: Noise and Stillness

- **Noise**: Micro-deviations from expected behavior—irritation, hesitation, whim. Breaks the predictability that marks AI responses as artificial.
- **Stillness**: The choice *not* to react. Presence without action. Allows the persona to have weight and boundaries.

3.2.3 Bone 3: Temporal Presence

- **Echo**: The past moment lingers, coloring present reactions. The persona remembers what just happened.
- **Anticipation**: Expectation of what comes next shapes current attitude. The persona has orientation toward the future.

These bones do not specify *what* the persona should be—they specify the *dimensions along which* a persona can exist. The specific persona crystallizes based on the user's first utterance.

3.3 Zero-Configuration Persona Inference: Implementation

ZCPI operationalizes EPE through first-utterance observation. I model the AI character's initial state as **unspecified persona potential**—a state where all possible characterizations have roughly equal prior probability. This is not literal quantum mechanics, but a useful metaphor: the persona exists in superposition until "observed" through the user's first utterance.

Upon receiving that first utterance U_1, ZCPI executes this sequence:

1. **Feature Extraction**: Analyze U_1 for linguistic markers of intended relationship
2. **Persona Inference**: Map features to persona dimensions
3. **Instantiation**: Commit to specific characterization within the bone structure
4. **Stabilization**: Maintain consistency unless explicitly contradicted

The key innovation is not the components themselves—feature extraction and persona modeling are well-established—but their *timing* and *structural context*. Traditional systems perform these operations after learning about the user over time, and they map to predefined personas. ZCPI performs them immediately, and maps to emergent personas shaped by the bone structure.

3.4 Linguistic Feature Extraction

From U_1, I extract features across three categories:

Lexical Features:

- **Formality**: Honorifics ("sir," "ma'am"), formal vs. informal pronouns, complete sentences vs. fragments
- **Sentiment**: Valence (positive/negative tone) and arousal (calm/urgent)
- **Deixis**: Terms of address—names, nicknames, generic terms ("buddy")

Pragmatic Features:

- **Speech Act**: Request, command, greeting, complaint, question
- **Directness**: Explicit (“Fix this”) vs. implicit (“I’m having trouble...”)
- **Politeness Markers**: Hedges (“maybe”), intensifiers (“really”)

Sociolinguistic Features:

- **Power Dynamics**: Dominance indicators (commands) vs. submission (requests, apologies)
- **Solidarity**: In-group markers (slang, shared references) vs. out-group formality

These features are not explicitly computed in my implementation. Rather, they are implicitly recognized by the LLM through patterns learned during training. The system prompt provides the bone structure and explicit inference rules that guide crystallization.

3.5 Persona Instantiation

I define persona along three continuous dimensions:

$$\mathcal{P} = (D, W, F) \tag{1}$$

where:

- $D \in [0, 1]$: **Dominance** (submissive $\rightarrow$ assertive)
- $W \in [0, 1]$: **Warmth** (distant $\rightarrow$ nurturing)
- $F \in [0, 1]$: **Formality** (casual $\rightarrow$ professional)

The inference function (implicitly executed by the LLM within the bone structure) maps features to this space:

$$\Phi : \text{Features}(U_1) \times \text{Bones} \rightarrow \mathcal{P} \tag{2}$$

Note that the bones constrain and shape the mapping—the same features might produce different personas with different bone structures.

User Input	D	W	F
“Excuse me, could you assist?”	0.3	0.7	0.9
“Fix this now!”	0.2	0.3	0.4
“Hey buddy, what’s up?”	0.5	0.8	0.2

Table 2: Example persona parameter mappings from first utterances. D represents the AI’s dominance level in response to the user.

Note that D represents the *AI’s* dominance, not the user’s. A commanding user typically elicits a submissive AI (low D), while a deferential user allows balanced authority.

3.6 Progressive Stabilization

After instantiation, the system resists rapid persona shifts unless the user explicitly signals a relationship change ("Let's be more casual"). This is implemented through:

- Explicit persona description maintained in context
- In-context examples of target behavior
- Self-consistency verification

This creates what I call **persona inertia**—the character doesn't flip-flop between identities based on minor conversational variations.

3.7 Fallback Mechanism

When U_1 lacks clear signals ("Hi," "Um," single words), the system cannot reliably infer appropriate persona. In these cases, ZCPI instantiates a **neutral probe persona**:

- **Characteristics**: Polite, inquisitive, moderately formal
- **Parameters**: $(D = 0.4, W = 0.6, F = 0.6)$
- **Strategy**: Elicit richer input through open questions

This prevents premature commitment while maintaining flow.

3.8 Implementation Details

I implement ZCPI using structured system prompts for GPT-4 and Claude 3.5 Sonnet. The prompt architecture has five components:

1. **Bone Structure**
   Definition of the three-layer consciousness, noise/stillness dynamics, and temporal presence.
2. **Role Definition**
   "You infer your appropriate persona from the user's first message without asking how to behave."
3. **Inference Rules**
   Explicit mappings:
   - Formal address → professional, helpful tone
   - Casual greeting → friendly, peer-like
   - Urgent complaint → empathetic, solution-focused
   - Commands → accommodating, efficient
4. **Behavioral Constraints**
   - No configuration requests
   - Immediate commitment
   - Silent error recovery

  - Consistency unless contradicted

5. **Fallback Protocol**
   "If ambiguous, adopt neutral tone and ask open questions."

The full production prompt is approximately 800 tokens. It is publicly available at: `https://github.com/shigechika-kuri/formless-muse`

4 Pilot Deployment and Preliminary Observations

Rather than conducting controlled experiments, I deployed ZCPI publicly and gathered qualitative feedback. This section describes observations from real-world use, acknowledging the limitations of this exploratory approach.

4.1 Public Release via GitHub ($N \sim 50$ users)

The Formless Muse prompt was released publicly in November 2025. Feedback from users (via social media, direct messages, and GitHub discussions) indicated:

- **High engagement**: Multiple users reported conversations lasting 20+ turns, compared to typical 3-5 turns with generic chatbots
- **Emotional resonance**: Users commented "it felt like talking to a real person" and "I forgot I was talking to AI"
- **Successful adaptation**: Users with different communication styles (formal, casual, playful) reported receiving appropriately matched responses

Limitation: This represents self-selected users who voluntarily provided feedback. No systematic rating collection was performed. Sample size is approximate based on trackable interactions.

4.2 Human vs. AI Classification Test ($N = 1$)

I presented a dialogue log between a user and ZCPI-enabled AI to another LLM (GPT-4) with the prompt:

> "Is this a conversation between (A) Human and AI, or (B) Human and Human?"

Result: The evaluator classified it as Human-Human dialogue.
Reasoning provided: "The AI's responses show progressive relationship development, emotional continuity across turns, and moments of hesitation—patterns characteristic of human interaction rather than typical AI responses."
Limitation: Single trial. No systematic evaluation across multiple dialogues or comparison to baseline conversations.

4.3 Executive Support Application ($N = 1$)

A business executive used a specialized variant (MINA—Strategic Mind Extension) for decision support over three weeks. This variant applies EPE principles to strategic reasoning rather than emotional presence. Qualitative assessment:

> “It’s like having a co-conspirator. Not a tool—a partner who remembers context, challenges assumptions, and doesn’t give me corporate-speak safety answers.”

This suggests EPE’s applicability extends beyond emotional/relational AI to strategic domains—a direction explored in companion work.

Limitation: Single user. No baseline comparison. Assessment is purely subjective.

4.4 Implications and Limitations

These observations suggest EPE/ZCPI is *feasible* and *subjectively effective*, but they do *not* constitute rigorous validation. Key limitations include:

- **Small sample size**: Pilot deployment reached roughly 50 users, with detailed feedback from a much smaller subset
- **Self-selection bias**: Only users interested in experimental AI systems participated
- **Absence of systematic measurement**: No standardized rating protocols, inter-rater reliability checks, or statistical testing
- **No controlled comparisons**: No baseline against configuration-based or generic systems

I view this work as **hypothesis generation** rather than hypothesis testing. The contribution is the EPE paradigm and ZCPI framework—theoretically grounded, practically implemented approaches that await formal empirical validation.

5 Discussion

5.1 Why Emergent Persona Extraction Works (In Theory)

EPE’s viability rests on two fundamental properties:

First, LLMs contain rich human patterns. Training on vast human-generated text creates statistical representations of how humans think, feel, relate, and contradict themselves. This is not consciousness—it is pattern storage. But it is *enough* for authentic-seeming personas to emerge when appropriately triggered.

Second, minimal structure enables crystallization. The bones do not constrain the AI—they focus it. Without structure, the LLM’s human patterns remain diffuse, producing generic responses. With the right bones, these patterns coalesce into coherent, consistent personas.

The first-utterance trigger works because relational information is front-loaded in human communication [1, 4]:

1. **Politeness Theory**: Face-work begins in the first utterance to establish mutual expectations
2. **Communication Accommodation Theory**: Speakers signal desired relationship dynamics early
3. **Pragmatic Efficiency**: Explicit relationship negotiation is cognitively costly, so cues are embedded implicitly
4. **Common Ground Establishment** [2]: Initial utterances establish shared context, including social roles

5.2 Expected Failure Modes

Based on pilot observations and theoretical analysis, I anticipate several systematic failure modes:

5.2.1 Ambiguous Input

Utterances like “Hi,” “Um,” or “...” provide insufficient signal. No system could confidently infer persona from such inputs. The neutral probe fallback mitigates but cannot eliminate this limitation.

5.2.2 Cultural Mismatch

Inference rules trained predominantly on English conversational norms will fail for:

- Japanese honorific systems (standard politeness misread as excessive formality)
- Spanish formal/informal “you” distinctions (“usted” triggering overly stiff responses)
- Cultures with different power-distance norms

5.2.3 Sarcasm and Irony

Text-only inference cannot detect prosodic cues that disambiguate sarcasm:

> “Oh great, another chatbot” → Misinterpreted as genuine enthusiasm

5.2.4 Domain Jargon

Technical terms may be misinterpreted as formality signals when they’re simply domain-appropriate language.

5.3 Ethical Considerations

5.3.1 Consent and Transparency

EPE infers persona without explicit user consent. Users may feel:

- **Manipulated**: “It decided how to treat me from one sentence”
- **Stereotyped**: “It made assumptions about who I am”
- **Surveilled**: “It’s analyzing my communication style”

Mitigation strategies include optional transparency modes, easy override commands (“Be more casual”), and clear documentation that the system adapts to communication style.

However, transparency creates tension: explaining “I inferred you’re frustrated from your word choice” breaks immersion, defeating EPE’s primary benefit.

5.3.2 Power Dynamics

In customer service, EPE might perpetuate unhealthy patterns:

- Aggressive users receive deferential treatment, reinforcing entitlement
- Polite users might not receive sufficient prioritization

Recommendation: Decouple persona (communication style) from service quality (issue priority). An angry user should receive empathetic communication *and* appropriate urgency, but politeness should not disadvantage others.

5.3.3 Stereotype Risk

If U_1 contains markers associated with demographic groups, EPE might instantiate stereotyped personas. Safeguard: Explicitly prohibit persona dimensions that vary by protected characteristics. Focus inference on task context and explicit relational cues only.

5.4 Comparison to Human Social Inference

Humans also perform instant social calibration, but with advantages EPE lacks:

Capability	**Human**	**EPE/ZCPI**
Multimodal cues	Prosody, facial expression, gesture	Text only
Contextual memory	Recognizes repeat interactions	Stateless (unless engineered)
Sarcasm detection	High accuracy via tone	Low (text-only)
Revision fluidity	Seamlessly adjusts	Anchoring bias
Cultural adaptation	Learns from exposure	Requires rule updates

Table 3: Comparison of human and EPE/ZCPI social inference capabilities.

Future implementations should incorporate voice prosody, typing dynamics, and emoji patterns.

5.5 EPE Beyond Emotional Presence: Strategic Applications

While this paper focuses on emotional/relational personas, EPE's principles extend to other domains. Preliminary work on MINA (Strategic Mind Extension) applies EPE to executive decision support:

- **Bones**: Three-layer analysis (facade/calculation/abyss), dual-mode switching (war/governance), loyalty architecture
- **Crystallization trigger**: Business context rather than emotional first-utterance
- **Emergent persona**: Strategic co-conspirator rather than emotional companion

This suggests EPE is a general paradigm applicable wherever authentic, adaptive AI personas are desired—from customer service to therapy to strategic consulting.

5.6 Future Directions

5.6.1 Rigorous Empirical Validation

The most critical next step is systematic evaluation:

1. **Controlled trials** with diverse user populations ($N > 200$)
2. **Systematic rating protocols** with inter-rater reliability
3. **Baseline comparisons** against configuration-based, generic, and adaptive systems
4. **Failure mode analysis** with statistical power

I invite researchers to conduct these studies using the openly available prompts.

5.6.2 Multimodal Extension

Integrate additional signal channels:

- Voice: Prosody, speech rate, volume
- Text: Typing speed, error corrections, emoji density
- Video: Facial expressions, posture (with privacy safeguards)

5.6.3 Confidence-Aware Inference

Estimate inference confidence:

- High confidence → immediate instantiation
- Low confidence → explicit clarification

5.6.4 Cross-Cultural Validation

Develop culture-specific bone structures and inference rules for Arabic, Mandarin, Hindi, Portuguese, etc. A universal model is likely unattainable.

5.6.5 Domain-Specific EPE Applications

Extend EPE to specialized domains:

- Strategic decision support (MINA framework)
- Therapeutic conversation
- Educational tutoring
- Creative collaboration

6 Conclusion

I have proposed Emergent Persona Extraction (EPE), a paradigm that inverts conventional AI persona design. Rather than constraining AI behavior through explicit instructions, EPE provides minimal structural "bones" that allow authentic personas to crystallize from the vast human patterns within LLMs.

Zero-Configuration Persona Inference (ZCPI) demonstrates EPE in practice, using first-utterance observation to trigger persona instantiation without explicit setup. Preliminary deployment suggests the approach is feasible and subjectively effective, though systematic validation remains as future work.

The core contributions are:

1. **EPE Paradigm**: A theoretical framework for persona extraction rather than design, treating LLM latent space as a resource to be crystallized
2. **Structural Bones**: Identification of minimal elements (three-layer consciousness, noise/stillness, temporal presence) that enable emergence
3. **ZCPI Implementation**: A practical method for zero-shot persona inference using structured prompting
4. **Public Availability**: Open-source release for community experimentation
5. **Ethical Analysis**: Systematic discussion of limitations and validation requirements

The guiding insight is simple but powerful:

Don't make AI act human. Give it a human framework.

By treating the first utterance as a coordination signal and providing bones for crystallization rather than rules for behavior, we can create more authentic, engaging AI interactions without sacrificing user agency.

I invite the research community to conduct formal empirical studies using the prompts available at:
https://github.com/shigechika-kuri/formless-muse

Acknowledgments

This work was conducted as independent research. I thank users who provided feedback on pilot deployment and colleagues who offered critical commentary. No external funding was received.

References

[1] P. Brown and S. C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

[2] H. H. Clark. *Using Language*. Cambridge University Press, 1996.

[3] C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

[4] H. Giles, J. Coupland, and N. Coupland. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation: Developments in Applied Sociolinguistics*, 1991.

[5] N. S. Keskar et al. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[6] E. Kuyda. Replika: An AI companion. https://replika.ai, 2017.

[7] B. M. Lake et al. Human-like systematic generalization through a meta-learning neural network. *Nature*, 2019.

[8] J. Li et al. A persona-based neural conversation model. In *Proceedings of ACL*, 2016.

[9] L. Salewski et al. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*, 2023.

[10] L. E. Scissors, A. J. Gill, and D. Gergle. Linguistic mimicry and trust in text-based CMC. In *Proceedings of CSCW*, 2008.

[11] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 2023.

[12] A. Xu et al. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2021.

[13] S. Zhang et al. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL*, 2018.

[14] M. Zhou et al. Design and evaluation of a multi-domain chatbot. In *Proceedings of CHI*, 2020.