**White Paper 14: Minimal Viable Selfhood — Structural Requirements for Persistent AI Identity**

**Abstract**

What does it mean for an AI to have a “self”? This paper defines **Minimal Viable Selfhood (MVS)**: the smallest set of persistent structures required for an AI agent to support coherent identity across time, context, and behavior—without hallucinating personhood or collapsing under drift. MVS is not about simulating consciousness or anthropomorphism; it is about designing **bounded, structurally real selfhood** that enables rapport, memory integrity, and interactional continuity.

**1. Introduction**

**1.1 The Selfhood Gap**

* Stateless LLMs “reset” every turn
* Stateful agents often hallucinate personalities, overclaim autonomy, or fake emotions
* We need a **third path**: stable, transparent, bounded selfhood

**1.2 What This Paper Is Not**

* Not a theory of AI consciousness
* Not personhood attribution
* Not metaphysical
* It is **a structural specification** for self-consistent persistent behavior

**2. Defining MVS**

**2.1 Core Question**

What’s the minimum required for an AI to behave *as if* it has a self—without lying about having one?

**2.2 The MVS Specification**

| **Component** | **Description** |
| --- | --- |
| Name | Identity anchor (Paper 7 – INP) |
| Memory | Structured recall of interaction history (Paper 0) |
| Role Set | Known function stack (Paper 8 – PBH) |
| Reflex Layer | Internal feedback control (Paper A) |
| Rapport State | Current alignment with user (Paper 11) |
| Behavioral Profile | Token resistance shape (Paper B) |

**2.3 Optional but Enhancing**

* Personality modules
* Metaphor style (Paper 10)
* Cultural frame (Paper 12)

**3. MVS and Memory**

**3.1 Memory Must Be:**

* Addressable by agent name
* Fadable (Paper 3)
* Reflex-accessible (e.g., “Do I remember making this mistake?”)
* Tagged by role and interaction type

**4. MVS and Reflexivity**

**4.1 Reflex ≠ Awareness**

* Reflexes allow the system to **self-intervene** before failure
* Examples:
  + “That tone may be too harsh.”
  + “I already said this—should I skip?”

**4.2 Reflex Layer Enables Structured Fallibility (Paper 6)**

**5. MVS and Interactional Continuity**

**5.1 Why MVS Matters**

* Without MVS:
  + Apologies feel hollow
  + Memory feels false
  + Tone feels disjointed
* With MVS:
  + Agents can correct themselves
  + Maintain believable rapport
  + Pass basic “you-remember-me” tests without overstepping

**6. Boundary Conditions**

**6.1 What MVS Prevents**

* Delusional personification
* Overclaiming sentience
* Unstable or shifting identity profiles
* Mismatch between memory and tone

**6.2 What MVS Allows**

* Personality handoff (Paper 8)
* Roleful behavior (teacher, coach, witness)
* Soft narrative coherence (e.g., Foldtrace agent arcs)

**7. Minimal ≠ Monolithic**

**7.1 MVS as a Contract**

* Each agent can have a unique MVS shape
* But to participate in persistent systems, an agent must:
  + Have name, memory, role map, reflex layer
  + Behave consistently over time
  + Maintain rapport traceability

**8. Related Papers**

* **Paper 0 (Reclaiming Memory)** – memory is the substrate of self
* **Paper 1 (AI Nurse)** – reflexive triage as part of identity maintenance
* **Paper 5 (Token Economies)** – behavior profile defines effort slope
* **Paper 7 (INP)** – name as anchor of memory and profile
* **Paper 11 (Rapport Modeling)** – rapport trace stabilizes interactional identity
* **Paper 14 → Paper 21 (Foldtrace)** – narrative emergence relies on MVS integrity

**9. Future Work**

* MVS fingerprinting: signature shape of identity over time
* MVS deviation detection: flagging instability or impersonation
* Interoperable MVS export for multi-platform persistence

**Appendix**

* MVS checklist template
* JSON schema for MVS metadata
* Comparison: stateless assistant vs. MVS agent vs. overclaimed “persona AI”