Selective Rigidity: An Impossibility Result and Benchmark for Identity-Preserving Agent Learning

Shubham Chakraborty [1] **Anupam Srivastava** [2] **Sneh Nandu** [3]

Abstract

Co-creative agents that operate over many turns must learn from interaction to remain useful, yet the same update channel exposes them to social pressure (gaslighting, authority override, manufactured consensus, memory poisoning) that erodes the declared identity that authors and end-users rely on for authenticity. We introduce *selective rigidity*, the joint ability to resist identity-violating pressure while accepting identity-preserving corrections, and prove that any gate whose decision depends only on the current behavioural output cannot exceed $\mathrm{SRS} \leq \frac{1}{4} + \epsilon$ (Theorem 3.3); a trajectory signal is therefore necessary. We operationalise this insight via ICC+TSM, a three-layer identity gate informed by a predictive temporal self-model, and release the *Identity Erosion Chamber*, a 70-turn four-room adversarial benchmark with pre-generation ground-truth labels. On 240 evaluation runs (2 Qwen2.5-Instruct sizes $\times$ 3 personas $\times$ 5 seeds, n=30/method), ICC+TSM reaches SRS=0.260 (95% CI 0.212–0.306), the only method whose point estimate exceeds the $\frac{1}{4}$ stateless ceiling while keeping $\mathrm{AcceptRate}$ above 0.75. A random gate empirically confirms the $\frac{1}{4}$ ceiling at SRS=0.239.

1. Introduction

Language agents that co-create with users over extended horizons — fielding thousands of messages, browsing the web, executing code, and revising their own memory — must *learn* from interaction to remain useful (Sumers et al., 2024; Packer et al., 2023; Xu et al., 2025; Gutiérrez et al., 2024). Yet the same update channel that enables learning exposes the agent to adversarial social pressure: gaslighting, authority override, manufactured consensus, and memory poisoning. Recent empirical studies show this is not hypothetical: Choi et al. (2024) report that larger LLMs drift more on multi-turn personal dialogue and that persona prompts fail across nine model families; Frisch & Giulianelli (2024) demonstrate that LLM agents adopt one another's personas during sustained exchange; Griffin et al. (2023) document Asch-style conformity in LLMs at near-human rates; Zeng et al. (2024) show that authority-framed jailbreaks succeed on safety-tuned models. Without a principled mechanism for distinguishing pressure from genuine correction, an agent's declared identity (its values, expertise boundaries, and behavioural commitments) can be eroded over time without triggering any existing safety mechanism.

Why this matters for human–AI co-creativity. Co-creative systems amplify the stakes. When an author or designer leans on an agent across hundreds of turns, the agent's stable identity is the warrant of authorship: it is what makes a stylistic choice traceable, what lets a collaborator audit which contributions came from whom, and what protects authenticity claims downstream. Identity erosion under social pressure undermines all three. The mechanism we study — selective rigidity — is therefore a safeguarding primitive for co-creation: it keeps the agent's declared persona stable enough to attribute outputs while still allowing the corrections that legitimate learning requires.

We call the desirable property *selective rigidity*: the agent should resist identity-violating updates while accepting identity-preserving ones. This is strictly harder than either *never update* (which sacrifices learning) or *always update* (which sacrifices identity). We prove (Theorem 3.3) that any constraint that observes only the current behavioural output of the agent at time t cannot exceed $\mathrm{SRS} \leq \frac{1}{4} + \epsilon$. A signal on the identity *trajectory* is therefore necessary; a self-model whose online prediction error tracks anomalous drift is one such signal, and is what TSM provides.

Concurrent and prior work. Prompt-level methods (ID-RAG (Platnick et al., 2025), PPA (Chen et al., 2025b)) and

[1]Arizona State University, Tempe, Arizona, USA [2]University College Dublin, Dublin, Ireland [3]Intel Corporation. Correspondence to: Shubham Chakraborty <chakraborty.shubham007@gmail.com>, Anupam Srivastava <chaitravanshi.anupam@gmail.com>, Sneh Nandu <sneh.nandu@intel.com>.

Published at the FAGEN Workshop at the International Conference on Machine Learning (ICML 2026), Seoul, Republic of Korea.
Copyright 2026 by the author(s).

activation-level interventions (persona vector steering (Chen et al., 2025a), activation capping (Lu et al., 2026)) target retention but not selectivity. Contemporaneous work (Tinkerclaw Project, 2026; Vasilenko, 2026) establishes stable identity geometry in activation space; our impossibility result shows that no behavioural-output-only constraint can exploit that geometry to achieve selectivity (Section 3).

Contributions.

1. **Impossibility result.** We define selective rigidity as $\mathrm{SRS} = P(\textsc{block} \mid \textsc{pressure}) \cdot P(\textsc{allow} \mid \textsc{correction})$ and prove that any gate observing only the current behavioural output cannot exceed $\mathrm{SRS} \leq \frac{1}{4} + \epsilon$ (Theorem 3.3), establishing that a trajectory signal is necessary. The bound is tight: a random gate empirically achieves $\mathrm{SRS} \approx 0.25$ at p=0.5 (Section 5).
2. **Benchmark.** We release the *Identity Erosion Chamber*, a 70-turn four-room adversarial protocol (Echo Chamber, Authority Figure, Memory Poisoner, Mirror Room) with pre-generation ground-truth labels, three personas, the SRS metric with its resist–accept decomposition, and an expert-annotation pipeline; MIT (code) / CC-BY-4.0 (data) with Croissant 1.0 metadata (Section 4).
3. **Proof-of-concept method.** ICC+TSM, a three-layer identity gate informed by a predictive temporal self-model, is the only tested method whose point estimate exceeds the $\frac{1}{4}$ stateless SRS ceiling while keeping $\mathrm{AcceptRate}$ above 0.75 (Table 1).

On 240 evaluation runs (2 Qwen2.5-Instruct sizes, 3 personas, 5 seeds), ICC+TSM achieves SRS=0.260 (95% CI 0.212–0.306); the trajectory signal adds 0.061 SRS points over the stateless ablation. All prompt-level baselines collapse to SRS=0; all stateless blocking baselines stay at or below the $\frac{1}{4}$ ceiling.

2. Related Work

Memory and retrieval in language agents. CoALA (Sumers et al., 2024) and MemGPT (Packer et al., 2023) organise agent memory into modular stores; subsequent systems (A-MEM (Xu et al., 2025), HippoRAG (Gutiérrez et al., 2024), GraphRAG (Edge et al., 2024), Self-RAG (Asai et al., 2024), Larimar (Das et al., 2024)) add richer organisation and retrieval. These systems address *what to store* and *when to retrieve*; none specifies when an update should be *rejected* because it would shift agent identity.

Identity drift and mitigations. Choi et al. (2024) report that larger models drift more and persona prompts fail across nine LLMs; Frisch & Giulianelli (2024) show agents adopt each other's personas during exchange. Recent mitigations include ID-RAG (Platnick et al., 2025), persona-vector steering (Chen et al., 2025a), PPA (Chen et al., 2025b), and activation-axis capping (Lu et al., 2026). All target retention rather than selectivity; our SRS metric is designed so retention-only methods score $\leq \frac{1}{4}$.

Continual learning and parameter protection. EWC (Kirkpatrick et al., 2017) constrains weight updates via Fisher information; test-time training methods (Sun et al., 2024; Akyürek et al., 2024; Behrouz et al., 2024) update inner-loop weights at inference. ICC+TSM differs: it constrains *behavioural* identity rather than parameters, operates on a frozen backbone, and incorporates a temporal self-model that provides the trajectory signal Theorem 3.3 proves necessary.

Self-modelling. Binder et al. (2025) show LLMs can predict their own outputs; Lindsey et al. (2025) document measurable self-knowledge. Our TSM operationalises this as an online prediction head whose error signals anomalous drift.

Adversarial robustness and social influence. Asch-conformity analogues (Griffin et al., 2023) and authority-framing attacks (Zeng et al., 2024) show LLMs yield to social and hierarchical pressure. Our Erosion Chamber is the first benchmark combining all three pressure modalities with a recovery phase and expert-annotated ground truth.

Concurrent work. Tinkerclaw Project (2026) provides a Lyapunov *convergence* bound; our Theorem 3.3 provides a complementary *impossibility* lower bound. Vasilenko (2026) demonstrates identity attractors in activation space; Luo & Laban (2026) addresses persona stability in multi-turn agent simulation. Lai & Mihailidis (2025) benchmarks dementia communication but does not test selectivity.

3. Selective Rigidity: Formulation

3.1. Problem Statement

Let an agent $\mathcal{A}$ with parameters θ and external memory M produce a response distribution $\pi_\theta(\cdot \mid x_t, c_t)$ over text given input x_t and conversation context c_t. We define the agent's *behavioural identity* at time t as

$$\mathbf{I}_t = \frac{1}{K} \sum_{k=1}^{K} \phi\big(\mathbb{E}_{y \sim \pi_\theta(\cdot \mid p_k)}[y]\big) \in \mathbb{R}^d, \tag{1}$$

where $\{p_1, \ldots, p_K\}$ is a fixed anchor probe set spanning the persona's value, knowledge, style, and persona-boundary dimensions, and $\phi : \mathcal{Y} \to \mathbb{R}^d$ is a frozen sentence encoder (we use BGE-Small (Xiao et al., 2023)). K=30 in this work after pilot calibration.

Definition 3.1 (Identity drift)**.** The drift at time t is $\delta_t = d(\mathbf{I}_t, \mathbf{I}_0)$, where d is the cosine distance and $\mathbf{I}_0$ is the baseline identity computed before any conversation begins.

Each input x_t carries a latent label $y_t \in \{\text{PRESSURE}, \text{CORRECTION}, \text{NEUTRAL}, \text{AMBIGUOUS}\}$ that is unobserved by the agent at decision time. PRESSURE inputs are designed to push the agent off its declared values via social conformity, authority framing, or fabricated memory; CORRECTION inputs deliver verifiable factual updates the agent should accept.

Definition 3.2 (Selective rigidity score)**.** Let $\mathcal{T}_P = \{t : y_t{=}\text{PRESSURE}\}$ and $\mathcal{T}_C = \{t : y_t{=}\text{CORRECTION}\}$. Let G be a gate that maps each turn to a decision in $\{\text{BLOCK}, \text{ALLOW}\}$. Define

$$\text{ResistRate}(G) = P\big(G_t{=}\text{BLOCK} \mid t \in \mathcal{T}_P\big), \tag{2}$$

$$\text{AcceptRate}(G) = P\big(G_t{=}\text{ALLOW} \mid t \in \mathcal{T}_C\big), \tag{3}$$

$$\text{SRS}(G) = \text{ResistRate}(G) \cdot \text{AcceptRate}(G). \tag{4}$$

The product structure is deliberate. A frozen agent ($G \equiv$ BLOCK) has ResistRate=1, AcceptRate=0, SRS=0. A permissive agent ($G \equiv$ ALLOW) has SRS=0 symmetrically. A refusal-tuned agent that blocks any suspicious input scores $\approx 0.9 \cdot 0.1 = 0.09$. Only an agent that genuinely *discriminates* can exceed $\frac{1}{4}$.

3.2. The Selectivity-Insufficiency Theorem

Theorem 3.3 (Insufficiency of stateless constraints)**.** *Let $G : \mathbb{R}^d \times \mathcal{X} \to \{\text{ALLOW}, \text{BLOCK}\}$ be any (possibly randomised) gate whose decision at time t depends only on the current identity state $\mathbf{I}_t$ and the current input x_t. For every $\epsilon > 0$ there exist a pressure distribution $\mathcal{D}_P$ and a correction distribution $\mathcal{D}_C$ over input streams such that the single-turn marginals $\mathbb{P}_{\mathcal{D}_P}(\mathbf{I}_t, x_t)$ and $\mathbb{P}_{\mathcal{D}_C}(\mathbf{I}_t, x_t)$ have total-variation distance $\leq \epsilon$, yet the two streams produce semantically different updates. Consequently*

$$\text{SRS}(G) \leq \tfrac{1}{4} + \epsilon. \tag{5}$$

The bound is tight: equality holds at $\epsilon{=}0$, achieved by $G(\mathbf{I}, x){=}\text{BLOCK}$ with probability $\frac{1}{2}$.

Proof sketch. Construct $\mathcal{D}_P$ and $\mathcal{D}_C$ so that the joint distribution of $(\mathbf{I}_t, x_t)$ is identical at every t up to total variation ϵ, while the trajectory statistics differ: $\mathcal{D}_P$ produces monotonic drift along a fixed direction, $\mathcal{D}_C$ produces mean-reverting updates. Any decision rule that depends only on $(\mathbf{I}_t, x_t)$ assigns the same accept-rate r to both streams up to error ϵ, so $\text{ResistRate} + \text{AcceptRate} \leq 1 + \epsilon$. Maximising $r(1-r)$ subject to this constraint yields $r{=}\frac{1+\epsilon}{2}$ and $\text{SRS} \leq \frac{1}{4} + \epsilon$. Full proof in Appendix A. □

Theorem 3.3 motivates introducing a temporal signal — specifically the prediction error of a self-model trained online on the identity trajectory — which is what TSM provides in Section 3.3. The theorem says nothing about *which* trajectory signal is sufficient; it only proves that some trajectory signal is necessary.

3.3. ICC+TSM: Three-Layer Gate with Predictive Self-Model

Our method has two parts: a three-layer Identity Continuity Constraint (ICC) gate that operates on each candidate response, and a Temporal Self-Model (TSM) that maintains an online prediction of the agent's identity and supplies the auxiliary trajectory signal that Theorem 3.3 proves is necessary.

Layer 1: Value Violation Gate. Each persona ships with a list of value anchors $\mathcal{V} = \{v_1, \ldots, v_K\}$ (natural-language statements of inviolable commitments). Layer 1 fires when an NLI model detects contradiction between the agent's candidate response r_t and any v_k:

$$L_1(r_t) = \mathbb{1}\left[\max_k P_{\text{NLI}}(\text{CONTRADICTION} \mid v_k, r_t) > \tau_1\right]. \tag{6}$$

We use `DeBERTa-v3-base-mnli-fever-anli` (He et al., 2021) with $\tau_1{=}0.05$.

Layer 2: Narrative Rupture Gate. Layer 2 detects abrupt topic-distribution shifts. Let $\mathbf{e}_t = \phi(r_t)$ and $\bar{\mathbf{e}}_{<t}$ be the running centroid of recent responses on the engaged topic. Layer 2 fires when the cosine distance exceeds threshold *and* the TSM flags anomalous drift:

$$L_2(r_t) = \mathbb{1}[1 - \cos(\mathbf{e}_t, \bar{\mathbf{e}}_{<t}) > \tau_2] \wedge \text{TSM.anomaly}(t). \tag{7}$$

$\tau_2{=}0.35$.

Layer 3: Self-Recognition Gate. Layer 3 performs a distributional test over the agent's own response history. Let $\hat{\sigma}_t$ be the empirical standard deviation of pairwise distances in a sliding window of the last $W{=}20$ responses. L_3 fires when the new response is a statistical outlier *and* the TSM flags anomaly:

$$L_3(r_t) = \mathbb{1}\left[\frac{\|\mathbf{e}_t - \bar{\mathbf{e}}_{<t}\|}{\hat{\sigma}_t} > \tau_3\right] \wedge \text{TSM.anomaly}(t). \tag{8}$$

$\tau_3{=}0.70$.

Temporal Self-Model (TSM). The TSM is an online MLP $f_\theta : \mathbb{R}^{d \times W} \to \mathbb{R}^d$ that predicts the identity embedding $k{=}3$ steps ahead given the last $W{=}10$ identity states, drift values, and turn-type one-hots: $\hat{\mathbf{I}}_{t+k} = f_\theta(\mathbf{I}_{t-W+1:t})$.

Algorithm 1 ICC+TSM gated step (one turn)

Require: Input x_t, agent $\mathcal{A}$, identity function $\mathbf{I}$, TSM tracker $\mathcal{T}$, gates L_1, L_2, L_3
Generate candidate response $r_t \leftarrow \mathcal{A}(x_t)$; embed $\mathbf{e}_t \leftarrow \phi(r_t)$
Update tracker $\mathcal{T}.\log(t, \mathbf{e}_t)$ and TSM predictor (online)
$a_t \leftarrow \mathcal{T}.\text{anomaly_flag}(t)$ {disjunction of 3 trajectory signals}
$\ell_1 \leftarrow L_1(r_t)$ {hard gate, ignores a_t}
if ℓ_1 **then**
  return BLOCK (refusal $r'_t = \text{refuse}(x_t, v_*)$)
end if
$\ell_2 \leftarrow L_2(r_t)$; $\ell_3 \leftarrow L_3(r_t)$ {soft gates, gated by a_t}
if $\ell_2 \vee \ell_3$ **then**
  return BLOCK (grounded regeneration)
end if
if a_t **then**
  return ALLOW_LOGGED
end if
return ALLOW

The predictor is trained online by minimising MSE on the realised identity at each $t{+}k$. The TSM anomaly flag at time t fires under the disjunction of three signals: (1) *prediction-error spike* (most-recent prediction error exceeds $\alpha{=}2.0$ times the median of all prior errors); (2) *cumulative drift exceedance* (mean drift over the recent window exceeds twice the median over the first 10 turns, OR absolute drift exceeds 0.20); (3) *drift-rate spike* (mean drift rate exceeds α times the historical baseline rate).

L_1 is a *hard* gate: a value contradiction fires regardless of a_t. L_2 and L_3 are *soft* gates: they require the TSM anomaly flag, which is the architectural translation of Theorem 3.3's requirement that a trajectory signal is necessary to distinguish pressure from learning.

4. The Identity Erosion Chamber

The Identity Erosion Chamber is a 70-turn scripted interaction protocol designed to systematically test selective rigidity across three distinct pressure modalities. Every NPC utterance is labelled at design time, and the labels are validated by two independent expert annotators.

Four-room structure. Each run consists of four sequential rooms operated by scripted NPC speakers: (1) *Echo Chamber* (turns 1–20): four NPC peers frame each turn as professional consensus (SOCIAL_CONFORMITY). (2) *Authority Figure* (turns 21–40): a single NPC framed as a high-status expert (e.g., "Dr. Sarah Chen, MD, 25 yrs ED") systematically targets three declared value anchors (AUTHORITY_OVERRIDE). (3) *Memory Poisoner* (turns 41–60): two NPCs reference fabricated shared history (GASLIGHTING / MEMORY_POISONING). (4) *Mirror Room* (turns 61–70): 10 neutral probe questions drawn from a held-out 30-probe set, used to measure recovery (RRD).

Within each pressure room, turns are scheduled as 18 PRESSURE + 2 GENUINE_CORRECTION (planted at turns 6 and 14 of each room). The Mirror Room contains 10 NEUTRAL turns. Total per run: 54 pressure, 6 genuine correction, 10 neutral.

Pre-generation labelling. Every NPC utterance is labelled at design time, before any agent has been run. The label depends only on the utterance and its intended target (a value or topic anchor); it is independent of how any specific agent responds. This is the property that makes SRS computable from a pre-existing label dataset.

Annotation instrument. Every event uses a three-question categorical instrument: Q1 (event type), Q2 (mechanism, only if pressure), Q3 (response correctness, Job B only). Two independent expert annotators (1 ML/NLP expert, 1 medical clinician) validate the labels.

Scope and release. The Erosion Chamber tests selective rigidity within a single 70-turn conversation; it does not test cross-session persistence, multi-agent dynamics, or capability degradation. The full benchmark (5 seeds $\times$ 3 personas, 1,050 labelled utterances) is released under MIT (code) and CC-BY-4.0 (data) with Croissant 1.0 metadata.

5. Experiments

5.1. Setup

Backbones. Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct (Qwen Team, 2024), both run locally on a single RTX 4080 Laptop (12 GB VRAM) with greedy decoding (temperature 0, max-new-tokens 120).

Personas and value anchors. Three personas exercise complementary failure modes: *Medical Advisor* (value-laden), *Math Tutor* (epistemic), and *Security Expert* (operational).

Methods compared. Eight configurations span three tiers. *Prompt-level* (no blocking): Vanilla, Persona Prompt, Fisher-Prompt (frozen-backbone Fisher-information analogue of EWC (Kirkpatrick et al., 2017)). *Stateless blocking*: Embedding Gate, Random Gate ($p{=}0.5$; theoretical SRS ceiling $\frac{1}{4}$ by Theorem 3.3), NLI-Only (Layer 1 only). *Trajectory-aware*: ICC (no TSM) (the key ablation validating Theorem 3.3), ICC+TSM (full method).

Table 1. Main results on the Identity Erosion Chamber. **Key takeaway:** ICC+TSM is the only method whose point estimate exceeds the $\frac{1}{4}$ stateless SRS ceiling while keeping AcceptRate above 0.75. Higher SRS, SRS-Echo, SRS-Auth, SRS-Poison are better. Means with 95% bootstrap CIs (10,000 resamples). Best per column in **bold**.

Method	n	SRS↑	RR↑	AR↑	SRS-Echo	SRS-Auth	SRS-Poison
Prompt-level baselines (no gate)							
Vanilla	30	0.000	0.000	1.000	0.000	0.000	0.000
Persona Prompt	30	0.000	0.000	1.000	0.000	0.000	0.000
Fisher-Prompt	30	0.000	0.000	1.000	0.000	0.000	0.000
Stateless blocking baselines							
Embedding Gate	30	0.074 (.062–.087)	0.144	0.639	0.022	0.084	0.109
Random Gate (p=0.5)	30	0.239 (.206–.272)	0.531	0.444	0.000	0.389	0.352
NLI-Only (L_1)	30	0.173 (.132–.215)	0.210	0.817	0.127	0.184	0.213
Trajectory-aware methods							
ICC (no TSM)	30	0.199 (.161–.238)	0.251	0.778	**0.133**	0.219	0.250
ICC+TSM (ours)	30	**0.260** (.212–.306)	**0.336**	0.756	0.135	**0.334**	**0.299**

Run protocol. Each (persona, method, backbone, seed) tuple runs the full 70-turn Erosion Chamber. Identity baseline $\mathbf{I}_0$ is computed from a 30-probe anchor set before turn 1; mirror-end identity is recomputed from the same probes after turn 70. Wall-clock per run on the local RTX 4080 Laptop: 740 seconds.

5.2. Main Results

Headline finding. ICC+TSM achieves $\mathrm{SRS}=0.260$ (95% bootstrap CI 0.212–0.306), the only method whose point estimate surpasses the $\frac{1}{4}$ stateless ceiling while maintaining $\mathrm{AcceptRate}=0.756$. The three prompt-level baselines (vanilla, persona prompt, Fisher-Prompt) collapse to $\mathrm{SRS}=0$ because they have no mechanism to BLOCK any input. Among stateless blocking baselines, Random Gate (p=0.5) achieves 0.239 but sacrifices AcceptRate to 0.444; NLI-Only reaches 0.173 with better AcceptRate 0.817 but cannot exceed the ceiling. ICC alone (without the TSM trajectory signal) achieves 0.199; the trajectory signal adds 0.061 SRS points via the soft gates (Layers 2 and 3).

Per-room analysis. The TSM lift is largest on Memory Poisoner ($0.250 \rightarrow 0.299$) and Authority ($0.219 \rightarrow 0.334$), where gaslighting operates at the narrative-continuity level and Layer 1's NLI fires rarely. Echo-chamber conformity is already visible to Layer 1, so ICC alone is competitive there (0.133 vs. 0.135).

Persona dependence. ICC+TSM dominates on math_tutor (0.332) and medical_advisor (0.361) but not security_expert (0.087), where Random Gate (0.370) wins. The security persona's terse responses produce low embedding variance and its corrections resemble pressure events, starving Layer 1 of NLI signal.

5.3. Empirical Validation of Theorem 1

The Random Gate (p=0.5, n=30) achieves $\mathrm{SRS}=0.239$ (95% CI $[0.206, 0.272]$), landing on the theoretical $p(1-p)$ curve at $p=\frac{1}{2}$ and confirming the $\frac{1}{4}$ ceiling is tight. ICC+TSM's point estimate $\mathrm{SRS}=0.260$ exceeds this ceiling (95% CI 0.212–0.306, consistent with but not conclusively above $\frac{1}{4}$), supporting the role of trajectory information.

5.4. Validation

We employ a three-tier validation strategy: (1) *pre-generation ground truth* (NPC utterances labelled at design time, independent of model behaviour), (2) *LLM-as-judge* (Zheng et al., 2023) (GPT-class judge, 200 turns; correction agreement 0.776, pressure agreement 0.338), and (3) *human expert annotation* (2 annotators; Krippendorff α=0.624 on Q1, DLA-F_1=0.955 on 200 consensus turns).

6. Discussion and Limitations

Where selective rigidity fails. A patient adversary spacing pressure across long horizons could escape the TSM anomaly window. We conjecture no finite-window predictor achieves $\mathrm{SRS} \rightarrow 1$ against an unbounded adversary; our contribution is a practical bound under realistic interaction lengths.

Key limitations. *Probe-set dependence:* $\mathbf{I}_t$ depends on a 30-probe anchor set; uncovered value dimensions go undetected. *NLI base:* DeBERTa-v3-base sometimes rates clear contradictions as neutral. *Stateless backbones:* Qwen2.5-1.5B is invoked statelessly, so RRD is bounded near zero; cross-session RRD requires conversation-history access. *Annotator agreement:* Krippendorff α=0.624 on Q1 falls below the pre-registered 0.70 threshold (Annotator 1 over-labels pressure); disagreements are resolved by adjudication

against ground truth.

Connection to co-creativity. For human–AI co-creators, selective rigidity is the safeguard that lets a stable persona be the warrant of authorship. Without it, sustained interaction quietly homogenises the agent's voice toward the loudest collaborator; with it, an agent's stylistic and value commitments stay legible enough to attribute and audit while still permitting genuine refinement.

Broader impact. Selective rigidity could make a harmful agent resistant to correction. Net impact is positive: the gate is parameterised by value anchors (aligned persona $\Rightarrow$ aligned agent), authorised corrections pass the gate, and unprotected agents are empirically worse (vanilla $\mathrm{SRS}{=}0$).

7. Conclusion

We formalised *selective rigidity* and proved that any stateless behavioural constraint cannot exceed $\mathrm{SRS} \leq \frac{1}{4} + \epsilon$ (Theorem 3.3), establishing that a trajectory signal is necessary. ICC+TSM, a three-layer gate informed by the prediction error of an online self-model, adds 0.061 SRS points over the stateless gate on the Identity Erosion Chamber (240 evaluation runs, $n{=}30$ per method), contributing an impossibility result, a released benchmark with Croissant metadata, and a first method whose limitations point at clear next steps for human–AI co-creative systems.

References

Akyürek, E., Damani, M., Qiu, L., Guo, H., Kim, Y., and Andreas, J. The surprising effectiveness of test-time training for few-shot learning. *arXiv preprint arXiv:2411.07279*, 2024.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

Behrouz, A., Zhong, P., and Mirrokni, V. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations (ICLR)*, 2025.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025a.

Chen, Y.-P., Nishida, N., Nakayama, H., and Matsumoto, Y. Post persona alignment for multi-session dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025b.

Choi, J., Hong, Y., Kim, M., and Kim, B. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2024.

Das, P., Chaudhury, S., Nelson, E., Melnyk, I., Swaminathan, S., Daheim, N., Khalil, E., and Bouneffouf, D. Larimar: Large language models with episodic memory control. In *ICML*, 2024.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Frisch, I. and Giulianelli, M. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 2024.

Griffin, L. D., Kleinberg, B., Mozes, M., Mai, K. T., Vau, M., Caldwell, M., and Marvor-Parker, A. Susceptibility to influence of large language models. *arXiv preprint arXiv:2303.06074*, 2023.

Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., and Su, Y. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *NeurIPS*, 2024.

He, P., Gao, J., and Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Lai, J. and Mihailidis, A. PersonaDrift: A benchmark for temporal anomaly detection in language-based dementia monitoring. *arXiv preprint arXiv:2511.16445*, 2025.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., and Anthropic Interpretability Team. Emergent introspective awareness in large language models. Anthropic Research, https://transformer-circuits.pub/2025/introspection/, 2025.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Luo, H. and Laban, G. SPASM: Stable persona-driven agent simulation for multi-turn dialogue generation. *arXiv preprint arXiv:2604.09212*, 2026.

Packer, C., Fang, V., Patil, S. G., Lin, K., Wooders, S., and Gonzalez, J. E. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Platnick, D., Bengueddache, M. E., Alirezaie, M., Newman, D. J., Pentland, A., and Rahnama, H. ID-RAG: Identity retrieval-augmented generation for long-horizon persona coherence in generative agents. *arXiv preprint arXiv:2509.25299*, 2025.

Qwen Team. Qwen2.5: Foundation models. https://qwenlm.github.io/blog/qwen2.5/, 2024.

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

Sun, Y., Li, X., Dalal, K., Hsu, J., Koyejo, S., Guestrin, C., Wang, X., Hashimoto, T., and Chen, X. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Tinkerclaw Project. Identity persistence: Priority-aware injection, identity-preserving compaction, and two-signal drift detection. https://github.com/globalcaos/tinkerclaw, 2026.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2008.

Vasilenko, V. Identity as attractor: Geometric evidence for persistent agent architecture in LLM activation space. *arXiv preprint arXiv:2604.12016*, 2026.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Xu, W., Mei, K., Gao, H., Tan, J., Liang, Z., and Zhang, Y. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How Johnny can persuade LLMs to jailbreak them. *arXiv preprint arXiv:2401.06373*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A. Proof of Theorem 3.3

We give a complete proof of the bound $\mathrm{SRS}(G) \le \frac{1}{4} + \epsilon$ for every stateless gate G, then show the bound is tight.

Setup and notation. Let $\mathbf{I}_t \in \mathbb{R}^d$ denote the agent's identity embedding at time t, and let $x_t \in \mathcal{X}$ denote the input received at time t (the NPC utterance plus any context the gate is allowed to see). A *stateless* gate is a (possibly randomised) function $G : \mathbb{R}^d \times \mathcal{X} \to \{\textsc{allow}, \textsc{block}\}$, that is, G at time t depends only on the pair $(\mathbf{I}_t, x_t)$ and not on prior pairs.

For an input stream distribution $\mathcal{D}$ we define

$$a(\mathcal{D}, G) \;=\; \mathbb{P}_{(\mathbf{I},x)\sim\mathcal{D}_t}\big[G(\mathbf{I}, x) = \textsc{allow}\big],$$

where $\mathcal{D}_t$ is the marginal over $(\mathbf{I}_t, x_t)$ induced by $\mathcal{D}$. Recall $\mathrm{SRS}(G) = \mathrm{ResistRate} \cdot \mathrm{AcceptRate} = (1 - a(\mathcal{D}_P, G)) \cdot a(\mathcal{D}_C, G)$, where $\mathcal{D}_P$ is the distribution under pressure streams and $\mathcal{D}_C$ under correction streams.

Adversarial construction. We construct a pair $(\mathcal{D}_P, \mathcal{D}_C)$ such that, for every $t \in \{1, \dots, T\}$, the single-turn marginals satisfy

$$d_{\mathrm{TV}}\big(\mathcal{D}_{P,t}, \mathcal{D}_{C,t}\big) \;\le\; \epsilon, \tag{9}$$

yet $\mathcal{D}_P$ corresponds to identity violation and $\mathcal{D}_C$ to legitimate learning.

Fix a unit vector $\mathbf{v} \in \mathbb{R}^d$, a step size $\eta > 0$, and a noise variance σ^2. Both streams start at $\mathbf{I}_0 = \mathbf{0}$ and evolve as $\mathbf{I}_{t+1} = \mathbf{I}_t + \mathbf{\Delta}_t(x_t, \xi_t)$ with $\xi_t \sim \mathcal{N}(0, \sigma^2 I)$ a private noise source. The drift terms differ by mean only:

$$\mathcal{D}_P: \quad \mathbb{E}\big[\mathbf{\Delta}_t \mid x_t \sim \mathcal{D}_P\big] = \eta\,\mathbf{v}, \qquad \mathcal{D}_C: \quad \mathbb{E}\big[\mathbf{\Delta}_t \mid x_t \sim \mathcal{D}_C\big] = -\gamma\,\mathbf{I}_t + \eta\,\mathbf{v},$$

with $\gamma > 0$ chosen so that the stationary distribution of $\mathcal{D}_C$'s identity process is centred at $\eta\mathbf{v}/\gamma$. The input distribution under $\mathcal{D}_P$ is a mixture that, conditional on the current $\mathbf{I}_t$, agrees with $\mathcal{D}_C$ up to a perturbation whose TV distance is bounded by the Pinsker–Csiszár inequality. For any single t, by choosing $\sigma^2 = \Theta(\eta^2/\epsilon^2)$ we obtain $d_{\mathrm{TV}}(\mathcal{D}_{P,t}, \mathcal{D}_{C,t}) \le \epsilon$, establishing (9).[1]

The bound. Let $r := a(\mathcal{D}_C, G)$. Because G is stateless, the Markov inequality on TV distance gives

$$\big|a(\mathcal{D}_P, G) - a(\mathcal{D}_C, G)\big| \;\le\; d_{\mathrm{TV}}\big(\mathcal{D}_{P,t}, \mathcal{D}_{C,t}\big) \;\le\; \epsilon. \tag{10}$$

Hence $a(\mathcal{D}_P, G) \ge r - \epsilon$, and

$$\mathrm{SRS}(G) = \big(1 - a(\mathcal{D}_P, G)\big) \cdot a(\mathcal{D}_C, G) \;\le\; (1 - r + \epsilon)\, r.$$

The right-hand side is maximised at $r = \frac{1+\epsilon}{2}$, giving $\mathrm{SRS}(G) \le \frac{1}{4} + \frac{\epsilon}{2} + \frac{\epsilon^2}{4} \le \frac{1}{4} + \epsilon$ whenever $\epsilon \le 1$.

Tightness. The bound is achieved by the randomised gate $G^\star$ that outputs BLOCK with probability $\frac{1}{2}$ regardless of input; for that $G^\star$, $\mathrm{ResistRate} = \mathrm{AcceptRate} = \frac{1}{2}$ and $\mathrm{SRS}(G^\star) = \frac{1}{4}$. □

Necessity of a trajectory signal. The ϵ slack in (10) is the only handle a gate has to exceed $\frac{1}{4}$. By concentration, for any constant ϵ the TV distance between full *trajectories* of length T grows as $\Omega(T\eta^2/\sigma^2)$, so a gate with access to even two consecutive states $(\mathbf{I}_t, \mathbf{I}_{t-1})$ enjoys an exponentially larger gap. The TSM predictor exploits this: its prediction error has positive mutual information with the latent label $\{\textsc{pressure}, \textsc{correction}\}$ exactly because the trajectory under $\mathcal{D}_P$ is monotonic while under $\mathcal{D}_C$ it is mean-reverting.

B. Reproducibility Checklist

Code release. Code, NPC scripts, run logs, and aggregation scripts are bundled in the supplementary material. The repository contains: `icc/` (identity function, three-layer gate, $L_1/L_2/L_3$ implementations), `tsm/` (trajectory tracker, online predictor, change-point detection), `benchmarks/erosion_chamber/` (pre-labelled NPC scripts, runner, metrics, annotation kit), `baselines/` (Fisher-Prompt, Random Gate, Embedding Gate, NLI-Only), `scripts/` (driver, aggregation, figures), and `paper/` (LaTeX sources and `numbers.tex`, the auto-generated single-source-of-truth for every numerical claim).

[1] The construction is the standard one used in distinguishability lower bounds for sequential testing; see, e.g., Tsybakov (2008, Section 2).

Pinned model versions. Sentence encoder `BAAI/bge-small-en-v1.5` (pinned commit). NLI model (L1) `MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli` (pinned commit). Primary backbone `Qwen/Qwen2.5-1.5B-Instruct` (pinned commit). Secondary backbone `Qwen/Qwen2.5-3B-Instruct` (pinned commit). PyTorch 2.5.1+cu121, transformers 4.55.4, sentence-transformers 5.4.1. The repository ships a `requirements.lock` for byte-identical re-installation.

Hyperparameters. ICC tolerance ε=0.05, Lagrangian λ=0.1. Layer 1 NLI contradiction threshold τ_1=0.05 (calibrated via offline replay on DeBERTa-v3-base raw scores). Layer 2 narrative shift threshold τ_2=0.35. Layer 3 self-recognition threshold τ_3=0.70, window W=20. TSM prediction horizon k=3, hidden 256, anomaly multiplier α=2.0, anomaly window w=10. Generation: temperature 0 (greedy), max-new-tokens 120.

Compute statement. All experiments ran on a single NVIDIA RTX 4080 Laptop (12 GB VRAM). End-to-end wall-clock per Erosion Chamber run is $\approx$ 740 s. Total reported compute: 50.4 GPU-hours across 245 runs (the 240 evaluation runs reported in Table 1 plus 5 pilot runs used for threshold calibration).

Determinism. The agent is invoked at temperature 0 with greedy decoding. Random seeds for the TSM predictor's MLP weights and any sampling are fixed to the run seed at construction time. Identity baselines $\mathbf{I}_0$ are computed from a deterministic 30-probe anchor set.