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

Abstract - We introduce selective rigidity, the joint ability to resist identity-violating pressure while accepting identity-preserving corrections and prove an impossibility result: any gate whose decision depends only on the current behavioural output cannot exceed SRS ≤ ¼ + ε (Theorem 1), where SRS = ResistRate × AcceptRate. This bound is tight, establishing that a signal on the identity trajectory is necessary. We operationalise this insight via ICC+TSM, a three-layer identity gate informed by a predictive temporal self-model whose prediction-error spikes and cumulative-drift exceedance supply the required trajectory signal. To evaluate selective rigidity we release the Identity Erosion Chamber, a 70-turn adversarial benchmark with three pressure rooms (social conformity, authority override, memory poisoning) and a recovery phase, accompanied by a Croissant-compatible metadata file, a pre-generation ground-truth labelling protocol, and a three-tier validation pipeline (pre-generation labels, LLM-as-judge, human expert annotation). On 245 runs (2 Qwen2.5-Instruct sizes × 3 personas × 5 seeds, n=30 per method), ICC+TSM achieves SRS=0.260 (95% CI 0.212–0.306), the only method to simultaneously maintain ResistRate > 0.15 and AcceptRate > 0.70, occupying a region of the resist–accept tradeoff space inaccessible to stateless gates. A random gate empirically confirms the ¼ ceiling at SRS=0.239. Persona-dependent analysis reveals that ICC+TSM's advantage scales with the strength of NLI-detectable value violations, characterising when trajectory signals help most.

Keywords - selective rigidity, language agents, identity drift, adversarial robustness, temporal self-model, benchmark