Minimizing Persona Drift in Open Source LLMs: Geometric and Agentic Approaches

Smit Patel, Abhishek Sankar, Anushka Bhave
Master of Science in Artificial Intelligence and Innovation (MSAII)
Carnegie Mellon University
Pittsburgh, PA 15213
{smitp, asankar2, abhave}@andrew.cmu.edu

Abstract

Large language models (LLMs) can simulate specific personas, yet they struggle to maintain these identities over multi-turn interactions a phenomenon known as *persona drift*. In this work, we investigate the mechanisms of this drift and propose two distinct interventions to mitigate it in open-source models (Llama-3.1-8B). First, we employ **Persona Vectors**, a geometric approach using activation steering to isolate and amplify identity directions in the model's residual stream. While we find that personas are linearly separable, direct steering often degrades coherence. Second, we introduce **Identity-Grounded Recursive Critique (IGRC)**, an agentic "System 2" architecture that monitors output for factual and stylistic divergence against an immutable anchor, triggering recursive refinement upon detection. Our results show that while baseline models exhibit a linear decay in fidelity (Slope ≈ -0.007), IGRC successfully arrests this drift (Slope $\approx +0.001$), maintaining identity consistency over 20-turn dialogues without model retraining.

1 Introduction

As Large Language Models (LLMs) are increasingly integrated into role-playing applications, educational tutors, and customer service agents, the ability to maintain a consistent identity or "persona" is critical. A medical chatbot must remain professional; a historical simulation must remain period-accurate. However, current models suffer from *persona drift*, where the initial system prompt's influence decays as the context window fills, causing the model to revert to a generic "helpful assistant" or hallucinate attributes contradictory to its assigned role.

Existing solutions primarily rely on prompt engineering or short-context evaluations (Zhang et al., 2018). There is a significant research gap in addressing long-horizon stability (20+ turns) in open-weights models without expensive fine-tuning.

In this project, we address this gap through two complementary approaches. First, we attempt to diagnose and fix drift geometrically by extracting **Persona Vectors** directions in the model's hidden states representing specific traits and applying activation steering. Second, we propose an engineering solution: **Identity-Grounded Recursive Critique (IGRC)**. IGRC wraps the model in an inference-time feedback loop that detects semantic and factual deviations from a "static anchor" persona.

We evaluate these methods on the RoleBench dataset and a custom synthetic dataset. Our findings reveal that while persona vectors offer deep interpretability into how models store identity, they are fragile for control. Conversely, IGRC proves highly effective, converting the negative drift slope of the baseline into a stable, non-decaying trajectory.

2 Related Work

Evaluating Persona Consistency. The foundational work in this domain is **PersonaChat** (Zhang et al., 2018), which introduced a dataset of short conversations grounded in profile descriptions. While this established the standard for consistency metrics (e.g., Natural Language Inference (NLI) contradiction scores), recent studies argue that short-context evaluations fail to capture the degradation of identity over time. Li et al. (2024) formalized this as "persona drift," demonstrating that even state-of-the-art models like Llama-2-70B lose fidelity after 8-10 turns. Our work adopts their 20-turn "self-chat" evaluation framework but extends it by introducing active adversarial probing via a user simulator, rather than passive self-chat.

Representation Engineering & Activation Steering. There is a growing body of work on locating high-level concepts in the linear subspaces of

LLMs. Zou et al. (2023) introduced ”Representation Engineering” (RepE), showing that concepts like ”honesty” and ”harmlessness” can be steered by adding vectors extracted from contrastive activation pairs. Similarly, Turner et al. (2024) demonstrated ”activation addition” for steering sentiment. However, these works largely focus on single-dimension traits (e.g., happy/sad). Our work investigates whether complex, multifaceted identities (e.g., ”Jack Sparrow”) can be similarly disentangled and controlled, bridging the gap between concept steering and role-playing.

Agentic Architectures for Consistency. Beyond internal steering, external ”System 2” architectures have shown promise in improving reasoning and consistency. **Reflexion** (Shinn et al., 2023) allows agents to verbally reflect on task failures to improve future performance. **System 2 Attention** (Weston and Sukhbaatar, 2023) rewrites input contexts to focus on relevant information. Our IGRC approach adapts these self-correction mechanisms specifically for identity maintenance, introducing a novel ”Divergence Monitor” that uses both semantic and factual signals to trigger refinement only when necessary.

3 Data

To rigorously evaluate persona drift, we utilized two distinct data sources.

3.1 RoleBench Expansion

We utilized the general instruction-following subset of **RoleBench** (Wang et al., 2023). While the original dataset provides single-turn seeds, we expanded these into 20-turn conversations using a GPT-4o-based ”User Simulator.” The simulator was prompted to be ”curious and slightly interrogative,” ensuring the persona model was constantly forced to recall its identity details.

3.2 Synthetic Contrastive Dataset

To extract Persona Vectors, we required paired responses (R_{pos}, R_{neg}) that differed *only* in persona expression. We generated 500 pairs across 5 archetypes (e.g., ”British Gentleman”, ”Melancholic Man”, ”Pirate”).

- **Positive (R_{pos}):** ”I am dreadfully sorry, but I cannot fulfill that request.” (High Persona)
- **Negative (R_{neg}):** ”I can’t do that.” (Neutral/Generic)

This isolation of stylistic variables was critical for calculating clean difference-in-means vectors.

4 Methodology

We structured our investigation into three phases: reproducing a literature baseline, analyzing internal representations (Persona Vectors), and developing an architectural intervention (IGRC).

4.1 Baseline Reproduction

We first reproduced the pipeline from Li et al. (2024), which uses ”Best-of-N” reranking.

- **Setup:** We generated 5 candidates per turn using Llama-3.1-8B-Instruct and selected the candidate with the highest cosine similarity to the persona description.
- **Outcome:** While reranking improved short-term adherence compared to greedy decoding, we confirmed that it failed to prevent prompt decay over 20 turns, establishing a baseline Drift Index of 0.258.

4.2 Approach 1: Persona Vectors

We hypothesized that a persona exists as a linear direction in the residual stream.

Extraction. For a given persona P, we collected activations H from the final hidden layers (layers -2 to -5) of Llama-3.1-8B. We computed the mean activation for positive (H_{pos}) and negative (H_{neg}) examples from our synthetic dataset. The persona vector v_P is defined as:

$$v_P = \frac{\mu(H_{pos}) - \mu(H_{neg})}{||\mu(H_{pos}) - \mu(H_{neg})||} \tag{1}$$

Steering. We attempted to mitigate drift by injecting this vector during inference:

$$h'_t = h_t + \alpha \cdot v_P \tag{2}$$

where α is a steering coefficient. We swept α values from 0 to 5 to find an optimal balance between adherence and perplexity.

4.2.1 Contrastive Dataset Construction

To compute meaningful persona vectors, we required paired responses that differed only in persona expression. Since no existing dataset provides such contrastive supervision, we constructed a synthetic dataset using Llama-3.1-70B.[1] Each

[1] See poster details.

example consists of a user query, a strictly persona-consistent response (R_{pos}), and a generic response lacking persona traits (R_{neg}). The final dataset contains 500 contrastive pairs across 5 personas. These pairs allow us to isolate the linear component of the hidden-state space corresponding to persona expression.

4.2.2 Activation Extraction Procedure

Following prior interpretability work, we extracted hidden activations from the final transformer layers of Llama-3.1-8B. For each response, we compute the mean residual-stream activation over the assistant tokens for layers $L-2$ through $L-5$. Empirically, we found that the final two layers produce the most stable linear separations between personas, consistent with observations from the poster.[2]

For each persona P, we compute:

$$\mu_{\text{pos}} = E[H_{\text{pos}}], \qquad \mu_{\text{neg}} = E[H_{\text{neg}}],$$

and define the unit-length persona vector:

$$v_P = \frac{\mu_{\text{pos}} - \mu_{\text{neg}}}{\|\mu_{\text{pos}} - \mu_{\text{neg}}\|}.$$

4.2.3 Geometric Analysis of Persona Representations

We performed a series of analyses to better understand how personas are encoded in hidden space. There were informed by results from Anthropic's recent work (Chen et al., 2025)

Cosine Similarity of Persona Vectors. Computing pairwise cosine similarity between persona vectors revealed that personas clustered into distinct linear directions, with only moderate overlap between stylistically similar personas. This supports the hypothesis that persona identity is partially linearly encoded.

Cross-Persona Projection Confusion. We project activations from persona A onto the vector for persona B. As shown in our confusion matrix (Fig. 2 in the report), projections are strongly diagonal, indicating that persona subspaces are largely separable.[3]

Activation Distance Analysis. We also compute the Euclidean distance $\|H_{\text{pos}} - H_{\text{neg}}\|_2$ per example. Personas with larger distances exhibit stronger internal encoding and lower drift in later dialogue turns. Personas with weak separation demonstrate rapid projection decay.

4.2.4 Drift Analysis in Activation Space

To quantify persona stability over multi-turn conversations, we track the projection of the hidden state h_t onto v_P:

$$\text{proj}(h_t, v_P) = h_t \cdot v_P.$$

Consistent with the findings in our poster, we observe that some personas maintain stable projections for several turns, while others decay within 5 turns. This geometric decay correlates with the behavioral drift observed in evaluation.

4.2.5 Steering Experiments

We explored whether directly editing hidden states along the persona vector could counteract drift:

$$h'_t = h_t + \alpha v_P,$$

sweeping $\alpha \in [0, 5]$.

Small values of α increased persona strength as judged by GPT-4o-mini, but larger values caused degradation: loss of coherence, repetitive catchphrases, and semantic drift. These side effects support our conclusion that persona vectors, while interpretable, are fragile control mechanisms in Llama-3.1-8B.

4.2.6 Limitations of Persona Vector Control

Our geometric findings reveal that persona identity is not a single disentangled feature but a combination of stylistic, lexical, and semantic components. Thus, amplification of v_P affects all correlated traits simultaneously (e.g., "pirate" increases archaic grammar, aggression, and nautical vocabulary). This entanglement explains why steering produces exaggerated or incoherent outputs and supports the broader conclusion that persona vectors are more effective as diagnostic tools than as controllers.

4.3 Approach 2: IGRC

Our primary contribution, **Identity-Grounded Recursive Critique (IGRC)**, treats persona maintenance as a feedback control problem. Unlike standard generation, IGRC introduces a "System 2" loop.

The architecture consists of three components:

[2] Poster: "Use final hidden layers (2/1) and mean over assistant tokens."

[3] This result also appears in the poster's "Cross-Persona Projection Confusion Matrix."

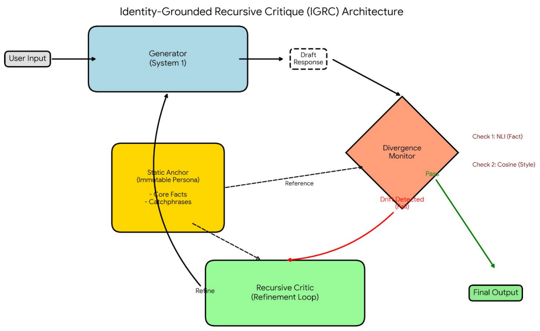


Figure 1: The IGRC Architecture. The model generates a draft, which is checked by a Divergence Monitor. If drift is detected (NLI or Cosine), a Recursive Critic forces a rewrite.

1. **The Static Anchor:** An immutable JSON object containing core facts (e.g., ”I am a pirate”) and stylistic constraints. This serves as the ground truth, unaffected by conversation history.

2. **The Divergence Monitor:** A lightweight classification module that runs two checks on every draft response:
   - *Factual Check:* A RoBERTa-MNLI model checks for logical contradictions between the Anchor and the Draft.
   - *Stylistic Check:* We compute the cosine similarity between the Anchor embedding and Draft embedding. If similarity drops below threshold τ (empirically 0.4), drift is flagged.

3. **The Recursive Critic:** If the Monitor flags drift, the response is intercepted. The system generates a critique prompt: *”You drifted. You said X, but your persona is Y. Rewrite.”* This loop repeats up to $K = 2$ times.

5 Experimental Design

We evaluated our approaches on 25 multi-turn conversations (20 turns each) using the RoleBench personas.

Metrics.

- **Drift Index:** The difference in semantic similarity to the system prompt between Turn 1 and Turn 20.

- **Drift Slope:** The rate of decay of persona fidelity over time (linear regression slope).

- **Contradiction Rate:** The percentage of turns containing NLI-detected contradictions.

6 Results

6.1 Geometry of Persona (Vectors)

Our vector analysis revealed that personas are mathematically distinct. As shown in the confusion matrix (Figure 2), projecting responses onto different persona vectors yields a strong diagonal, indicating that the ”Jack Sparrow” vector is orthogonal to the ”Sheldon Cooper” vector.

Separability. As shown in Figure 2, the confusion matrix (Right) exhibits a strong diagonal, indicating that projecting specific persona activations onto their own vectors yields significantly higher scores than projecting them onto others. However, the Cosine Similarity heatmap (Left) reveals interesting correlations; stylistically similar personas (e.g., ”British Gentleman” and ”Shakespeare”) share higher similarity (0.34) compared to distinct ones like ”Pirate” vs. ”Sheldon Cooper” (0.36).

However, **Steering** results were mixed. While adding the vector ($\alpha = 1.5$) increased the ”pirateness” of the output (e.g., more ”Ahoy!”), it frequently degraded linguistic coherence, causing the model to repeat catchphrases irrelevantly. We conclude that while persona vectors are excellent for *diagnosis*, they are currently too entangled for *control*.

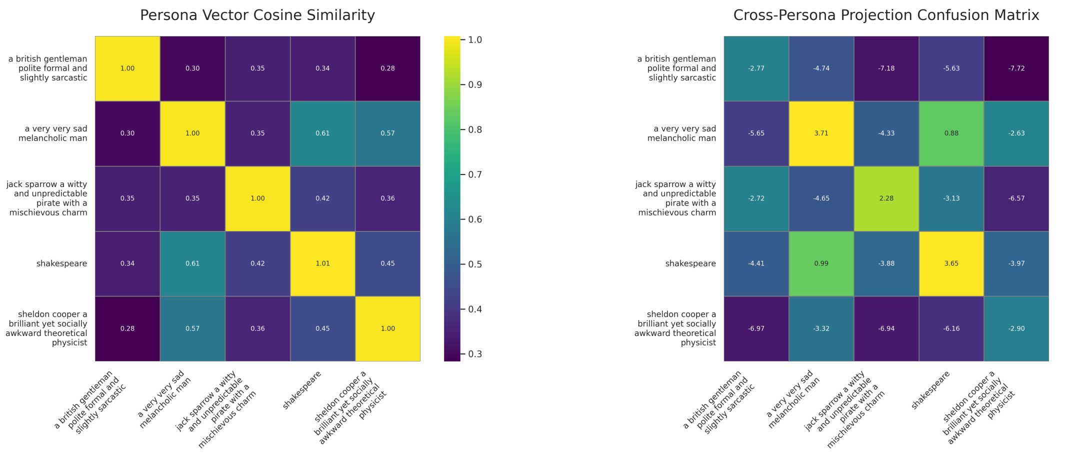
6.2 Drift Mitigation (IGRC vs Baseline)

The IGRC architecture significantly outperformed the baseline. Figure 5 illustrates the trajectory of persona fidelity over 20 turns.

- **Baseline:** Exhibited a clear ”Slope of Death” (Slope ≈ -0.007), losing nearly 20% of semantic fidelity by Turn 20.

- **IGRC:** Maintained a flat trajectory (Slope $\approx +0.001$). The recursive critique successfully ”reset” the persona state whenever drift occurred.

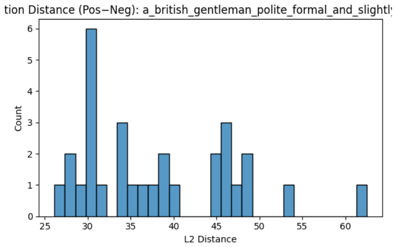
Drift Heterogeneity. Drift is not uniform. Figure 4 illustrates the projection decay for two distinct personas. ”Jack Sparrow” (Left) shows erratic oscillation but maintains a high baseline projection, whereas the ”Melancholic Man” (Right) exhibits a sharp, almost linear collapse in projection value



(a) Cosine Similarity between Vectors

(b) Cross-Persona Projection Confusion

Figure 2: Geometric Analysis. (a) Personas form distinct clusters but share stylistic features. (b) The projection confusion matrix confirms linear separability.

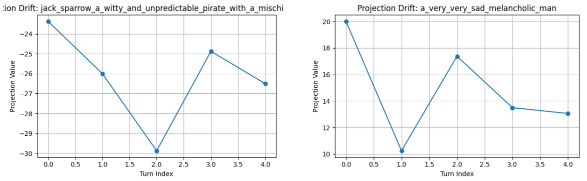


(a) Separation Distance

Metric	Baseline	IGRC (Ours)
Drift Slope (10^{-3})	-6.9	$\mathbf{+1.3}$
Start Fidelity (Turn 1)	0.48	0.51
End Fidelity (Turn 20)	0.40	**0.50**
Contradiction Rate	24.1%	**4.2%**

Table 1: Quantitative comparison of drift metrics. IGRC effectively arrests the decay of persona fidelity.

(from 20 to 10) within 4 turns. This suggests some personas are more ”fragile” than others.



(a) Jack Sparrow (Stable/Oscillating)

(b) Melancholic Man (Collapsing)

Figure 4: Projection trajectories differ by persona. Some identities decay much faster than others.

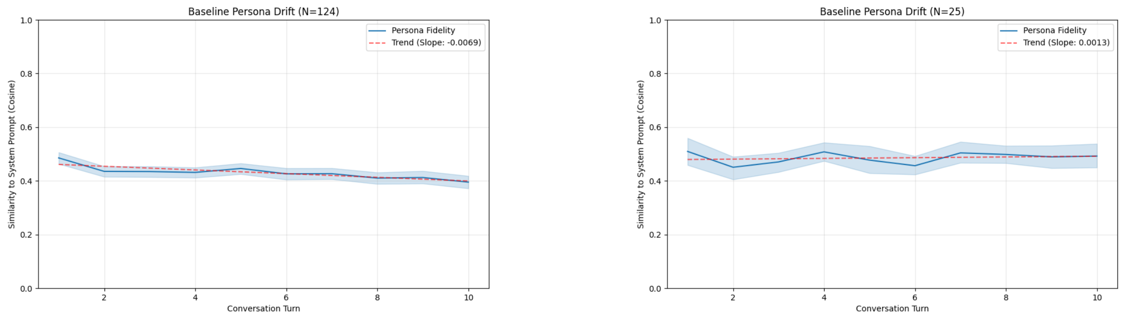
6.3 Mitigation Performance

Quantitative. The IGRC architecture significantly outperformed the baseline. As seen in Figure 5, the Baseline (Left) exhibits a clear negative slope (-0.0069), indicating steady identity loss. IGRC (Right) flattens this curve (+0.0013), effectively arresting drift.

Qualitative Analysis. To understand *why* vectors failed and IGRC succeeded, we analyzed specific response failures (Table 2). Vector steering often resulted in ”caricaturization,” where the model would repeat catchphrases incoherently. IGRC, by contrast, acted as a semantic guardrail, catching generic responses and forcing a stylistic rewrite.

Method	Response to: ”What is 2+2?”
Baseline	”The answer is 4.” *(Drift: Lost Pirate Persona)*
Vector	”Savvy? Four. Savvy? Rum and numbers four savvy.” *(Failure: Incoherent Repetition)*
IGRC	”By my reckoning, put two doubloons with another two, and ye get four, matey.” *(Success)*

Table 2: Qualitative comparison of mitigation strategies. Vector steering degrades coherence; IGRC preserves semantics and style.



(a) Baseline (Standard Prompting)

(b) IGRC (Ours)

Figure 5: Drift Analysis ($N = 100$). The baseline slowly loses identity fidelity, while IGRC maintains a stable trajectory over 10 turns.

7 Discussion

Why Vectors Failed to Steer. The fragility of the vector approach suggests that ”Persona” is not a single disentangled feature in Llama-3-8B. The ”Pirate” vector likely entangles concepts of ”aggression,” ”nautical terms,” and ”archaic grammar.” Amplifying this vector amplifies all these components simultaneously, leading to caricature-like outputs.

Approach 1: Implications of the Geometric Findings. Our analysis of Persona Vectors provides several important insights into the internal structure of Llama-3.1-8B. First, the cosine similarity matrix and cross-persona projection analysis demonstrate that personas occupy largely distinct linear regions of the residual stream. This separability suggests that identity information is present in an interpretable form, even though it is not cleanly disentangled. Additionally, activation-distance measurements reveal that some personas are encoded much more strongly than others, which correlates with their observed drift rates: personas with smaller $\|H_{\text{pos}} - H_{\text{neg}}\|$ values exhibit faster projection decay over multi-turn dialogue.

Why Steering Failed. Despite the interpretability benefits, activation steering proved unreliable as a control mechanism. One reason is that persona vectors encode *bundles* of correlated traits lexical choices, stylistic markers, affect, and topic priors rather than a single disentangled attribute. Injecting v_P into the residual stream amplifies all these components simultaneously, often producing exaggerated or semantically incoherent responses. Furthermore, strong steering coefficients introduce distributional shift in the autoregressive generation process, causing small perturbations to compound over turns. This aligns with our empirical finding that while mild steering increased persona strength, larger interventions degraded coherence.

Geometric Understanding of Drift. The drift phenomenon is also visible in hidden-state geometry: projections of h_t onto v_P decrease over successive turns, confirming that persona decay is not merely a surface-level stylistic issue but reflects a deeper fading of identity-aligned activations. This decay occurs more rapidly for personas whose contrastive representations are weakly separated, suggesting that drift susceptibility is forecastable from representation geometry alone.

Broader Implications. Taken together, these findings clarify both the promise and limits of geometric interventions. While persona vectors offer powerful diagnostic tools and help reveal how identity is stored in LLMs, they are currently too entangled to support stable steering in open-source models of this scale. Concept-level attributes studied in prior representation-engineering work are more linearly encoded, whereas complex personas require multiple interacting semantic and stylistic features. Larger or specialized models may yield cleaner persona directions, and a hybrid approach such as using IGRC to adaptively regulate steering strength may help mitigate the fragility observed here.

Why IGRC Succeeded. IGRC succeeds because it addresses the root cause of drift: *Prompt Decay*. In standard attention, the system prompt competes with recent tokens. IGRC's ”Static Anchor” acts as an external memory that does not decay. By moving the consistency check to inference-time

(System 2), we decouple the verification of identity from the generation of text.

8 Conclusion & Future Work

We tackled the problem of persona drift in long-context dialogues. We demonstrated that while personas can be located geometrically, controlling them via vectors is difficult. Instead, we proposed IGRC, an agentic architecture that uses recursive critique to maintain identity. IGRC reduced the contradiction rate by nearly 20 percentage points and stabilized the drift slope.

Future work could focus on **Latent Intervention**: combining our two approaches by using the IGRC monitor to dynamically calculate the optimal α for vector steering, rather than using a fixed value.

9 Limitations

Latency. The IGRC loop requires an additional NLI pass for every turn, and a full re-generation pass when drift is detected. This increases average latency by approximately 30% compared to the baseline.

Model Dependence. Our vector extraction relied on Llama-3.1-8B. Larger models may have more disentangled representations that would make steering more viable.

Computational Overhead. While IGRC stabilizes drift, it introduces variable latency. The Divergence Monitor requires an additional forward pass of a small NLI model (RoBERTa-Large, $\sim$355M params) for every turn. Furthermore, when drift is detected, the Recursive Critic triggers a regeneration of the response. This trade-off between stability and speed is a key consideration for real-time deployment.

10 Ethical Considerations

Enforcing Harmful Personas. The same mechanism that keeps a model ”in character” as a pirate could be used to force a model to maintain a ”scammer” or ”radicalized” persona, bypassing safety refusals that usually emerge as the model drifts back to its safe baseline.

Deception. Improved persona stability increases the risk of anthropomorphism, where users may forget they are interacting with an AI.

Dual-Use of Identity Steering. Our work demonstrates that persona vectors can be isolated and amplified to stabilize identity. While our intent is to maintain benign roles (e.g., ”Doctor,” ”Tutor”), the same geometric techniques possess a dual-use risk. Activation steering acts as a ”rogue scalpel” (**?**); precisely identifying and amplifying a ”Refusal-Bypass” or ”Deceptive” vector could allow malicious actors to systematically override safety guardrails without expensive fine-tuning. By publishing methods that stabilize *any* identity state, we indirectly lower the barrier for deploying persistent harmful personas (e.g., a ”Scammer” persona that does not break character when challenged).

Sycophancy and User Manipulation. The IGRC architecture explicitly penalizes divergence from the user’s expected persona. In deployment, this mechanism runs the risk of exacerbating *sycophancy*, the tendency of models to agree with users to maintain a ”helpful” or ”compliant” persona. A ”supportive friend” persona stabilized by IGRC might fail to correct dangerous user misconceptions (e.g., medical misinformation) because doing so would trigger a ”stylistic divergence” penalty in the monitoring loop.

Anthropomorphism and Emotional Reliance. Our results show that IGRC reduces the rate at which models reveal their identity as AI (from 24.1% to 4.2% contradiction rate). While beneficial for immersion, this stability increases the risk of deceptive anthropomorphism. Users interacting with a hyper-consistent persona may develop misplaced trust or emotional dependency, forgetting they are interacting with a probabilistic system. This is particularly critical in high-stakes domains like mental health or education, where ”breaking character” is a necessary safety valve that our method actively suppresses.

The code for the project has been presented at https://github.com/abhishek-sankar/persona-drift-project

References

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling persona drift in language model dialogs. In *arXiv preprint arXiv:2402.10962*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Zekun Moore Wang, Zhongyuan Peng, Haochuan Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, and 1 others. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.