Beyond Static Persona Consistency: Dynamic Persona Coherence in LLM Role-Playing

Yirui Qi[1]*, **Xiaoming Zhang**[1]†, **Ruilin Zeng**[1], **Mengyao Liu**[1], **Ziyi Zhou**[1], **Dezhuang Miao**[1], **Bingyu Yan**[1], **Zhenyu Guan**[1]

[1]Beihang University

Correspondence: yolixs@buaa.edu.cn

Abstract

Current LLM role-playing systems model persona as a monolithic, static attribute, conflating identity consistency with emotional rigidity. This leads to either robotic repetition or catastrophic persona drift under sustained interaction. We introduce Dynamic Persona Coherence, a framework that decouples Identity-Layer Stability (time-invariant traits) from Adaptive-Layer Appropriateness (history-dependent psychological evolution). We operationalize this through the L/M/S Psychological State Model, which represents persona dynamics across long-term identity, mid-term meaning/stress accumulation, and short-term affect. On top of this state representation, a closed-loop alignment system comprising an automated evaluator (Persona Consistency Critic, PCC), a selective repository (Persona Case Repository, PCR), and a trajectory-adjusting corrector (Persona Drift Suppressor, PDS) enables autonomous coherence repair. Experiments on GPT-4o, Claude-3.5-Sonnet, and DeepSeek-V3.2 demonstrate consistent improvements (+16–84% PCC gains). Code is available at `https://anonymous.4open.science/r/DPC-30A1`.

1 Introduction

Large language models (LLMs) have emerged as a transformative milestone in artificial intelligence (Achiam et al., 2023; Touvron et al., 2023). demonstrating unprecedented capabilities in natural language understanding and generation across diverse domains (Zhao et al., 2023; Naveed et al., 2025). LLMs have fundamentally reshaped downstream applications, particularly in role-playing settings where personas are simulated through conditioned text generation. Accordingly, recent research has increasingly focused on sophisticated persona simulation to achieve more authentic character portrayal (Zhou et al., 2025; Park et al., 2023).

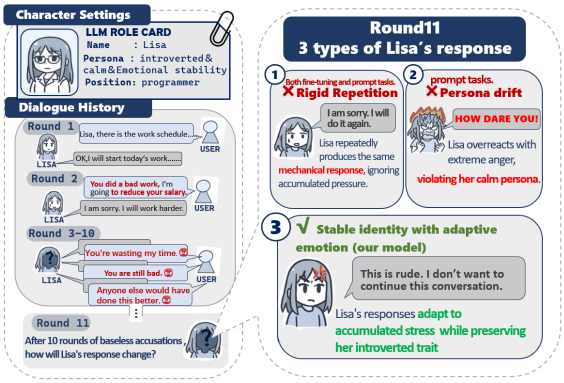

Figure 1: Persona coherence under sustained provocation. Existing models exhibit either robotic repetition or persona drift, whereas our approach enables lawful emotional adaptation, shifting from politeness to cold withdrawal while preserving a stable introverted identity.

To achieve this goal, recent researches have primarily adopted two strategies. Prompt engineering methods (Tseng et al., 2024) inject character descriptions at runtime, offering greater flexibility, which can elicit convincing behavior in single turns or short sessions but suffer from persona drift in extended interactions (Li et al., 2024; Frisch and Giulianelli, 2024). Fine-tuning approaches (Shao et al., 2023; Wang et al., 2024; Lu et al., 2024) fine-tune base models on constructed role-specific datasets to internalize character knowledge and linguistic styles, achieving notable improvements in character fidelity.

However, both approaches share a fundamental limitation: they model persona as a monolithic, static attribute. This limitation stems from the inherent nature of LLMs as disembodied neural networks trained on text corpora with next-token prediction objectives (Shanahan et al., 2023), which tend to represent persona as fixed configurations (whether through static prompt descriptions or

* First author.
† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28942–28956

frozen parameter encodings) rather than evolving psychological states.

Yet real individuals maintain stable trait baselines while exhibiting varying affective dynamics driven by accumulated psychological states (Fleeson, 2001). Current approaches conflate persona consistency (who they are) with emotional rigidity (how they react), failing to model how accumulated psychological states should shape lawful emotional evolution. As illustrated in Figure 1, when a"introverted, calm and emotional stability"persona faces sustained provocation, existing models either produce robotic repetition (ignoring cumulative stress) or catastrophic drift (breaking character entirely).In contrast, our approach enables coherent state evolution: as stress mounts, the persona transitions from politeness to cold withdrawal—adaptively escalating emotional intensity while strictly adhering to her introverted constraints.

To bridge this architectural disconnect, we propose a paradigm shift from static consistency to Dynamic Persona Coherence. Unlike prior definitions that treat persona as a monolithic attribute evaluated turn-independently, we formalize Dynamic Persona Coherence as a hierarchical constraint problem requiring: (i) Identity-Layer Stability (stable traits), and (ii) Adaptive-Layer Appropriateness (lawful state evolution under identity-constrained expression). Under this framework, a coherent agent must not only maintain stable identity anchors but also model history-dependent psychological evolution—tracking how accumulated experiences shape current state trajectories while ensuring adaptive variations are expressed through identity-consistent behavioral patterns. In summary, this paper makes the following contributions:

- Task Formalization: We introduce Dynamic Persona Coherence, the first framework (to our knowledge) that decouples Identity Stability from Adaptive Appropriateness, reframing persona consistency as a constrained state evolution process rather than turn-independent trait matching.
- Psychological Modeling: We propose the L/M/S (Long/Mid/Short-term) Psychological State Model, which disentangles immutable traits, accumulated stress, and transient affect, enabling multi-timescale psychological state tracking.
- Closed-loop Alignment: We develop a self-correcting system combining an automated evaluator (PCC) with a trajectory-adjusting simulator (PDS), enabling agents to autonomously detect and repair coherence violations without external supervision.

Experiments on GPT-4o (Achiam et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), and DeepSeek-V3.2 (Liu et al., 2024) show consistent gains, with +26.8% average PCC improvement over static baselines, particularly under adversarial stress-tracking scenarios.

2 Related Work

LLM Role-Playing Methods. LLM role-playing methods can be broadly categorized into prompt-based and fine-tuning approaches (Tseng et al., 2024; Nguyen et al., 2024). Early prompt-based work conditions LLMs at inference time using static persona cards (Zhang et al., 2018). Subsequent systems augment this paradigm with memory and retrieval modules (Li et al., 2023): Generative Agents (Park et al., 2023) integrate episodic memory, planning, and reflection to produce believable agent behaviors in simulated environments. Fine-tuning approaches internalize persona characteristics into model parameters (Zhou et al., 2024). Character-LLM (Shao et al., 2023) trains agents via experience reconstruction, achieving notable improvements on persona knowledge tasks. RoleLLM (Wang et al., 2024) constructs detailed profiles for 100 characters and applies role-conditioned instruction tuning to enhance fidelity. Despite these advances, both paradigms encode persona as static representations (whether in prompts or parameters) and fail to capture affective evolution under diverse event patterns (e.g., repeated success, alternating feedback, cumulative stress).

Evaluation, Theory, and Positioning. Evaluation of role-playing agents has evolved from reference-based metrics (e.g., ROUGE, BLEU) to multi-dimensional assessments (Tu et al., 2024; Samuel et al., 2025). Recent work adopts LLM-as-judge protocols for scalable assessment (Shao et al., 2023), and Shin et al. (Shin et al., 2025) propose atomic-level detection of out-of-character behaviors. However, these benchmarks predominantly assess fidelity to static profiles through monolithic scoring, conflating identity violations with state misalignment and lacking the decoupled evaluation necessary for diagnosing specific failure modes. This gap motivates grounding persona

modeling in established psychological theory. The theoretical foundation for separating stable identity from dynamic states is well-established in personality psychology. Trait–state theories (Mischel and Shoda, 1995) distinguish between enduring dispositions and transient emotional states, with cognitive appraisal theories (Smith and Lazarus, 1990; Scherer, 2001) explaining how traits bias event interpretations. Recent computational studies explore whether LLMs exhibit appraisal-like reasoning in emotional contexts (Tak and Gratch, 2023), though typically at a per-turn level without cumulative modeling.

3 Dynamic Persona Coherence Framework

3.1 Problem Analysis

Current persona consistency definitions suffer from two fundamental limitations: conflating stable personality traits with transient emotional states, and evaluating responses turn-independently without modeling cumulative psychological dynamics. We formalize an alternative criterion—**Dynamic Persona Coherence**—that addresses both via explicit trait-state decoupling and history-dependent state evolution.

Definition (Dynamic Persona Coherence). An agent exhibits dynamic persona coherence if it maintains *identity-layer stability* ($\mathcal{I}$: time-invariant traits defining "who" the character is) while exhibiting *adaptive-layer appropriateness* ($\mathcal{A}_t$:time-varying psychological states reflecting "how" the character feels at turn t), where adaptive states evolve cumulatively and are expressed through identity-constrained behavioral patterns.

3.2 Formal Framework

Let $\mathcal{I}$ denote the identity layer, $\mathcal{A}_t$ the adaptive state at turn t, E_t the input event, and R_t the generated response. An agent satisfies dynamic persona coherence over T turns iff:

$\mathcal{I}$-Invariance. The identity layer remains constant across all turns:

$$\forall t, t', \quad \mathcal{I}_t = \mathcal{I}_{t'} = \mathcal{I} \tag{1}$$

$\mathcal{A}$-Evolution. Adaptive states update as a function of prior state, current event, and identity:

$$\mathcal{A}_t = \text{Transition}(\mathcal{A}_{t-1}, E_t, \mathcal{I}) \tag{2}$$

Constrained Expression. Responses must preserve identity while reflecting current state:

$$R_t \in \text{ValidSet}(\mathcal{I}) \tag{3}$$

where $\text{ValidSet}(\mathcal{I})$ denotes all responses consistent with identity $\mathcal{I}$.

3.3 Operationalization

To empirically verify dynamic coherence, we introduce two complementary evaluation principles:

M1: Decoupled Scoring. Rather than conflating traits and states in a single score, we evaluate $\mathcal{I}$-stability and $\mathcal{A}$-appropriateness separately, then aggregate via:

$$\text{coherence} = \min(\mathcal{I}_{\text{score}}, \mathcal{A}_{\text{score}}) \tag{4}$$

The minimum operator ensures neither dimension can be sacrificed for the other.

M2: Temporal Tracking. Instead of scoring turns in isolation, we examine trajectory patterns: (i) $\mathcal{I}$-stability requires low variance across turns ($\text{Var}_t(\mathcal{I}_{\text{score}}) < \epsilon$), and (ii) $\mathcal{A}$-evolution must follow lawful transitions rather than random jumps. §4.4.1 details how PCC implements both principles.

3.4 Implementation Challenges

This framework poses three challenges:

- **C1: Partition and Transition.** How to partition $\mathcal{I}$ vs. $\mathcal{A}$ and instantiate Transition? → §4.3 proposes a three-layer (L/M/S) psychological model where L instantiates $\mathcal{I}$ and S/M jointly realize $\mathcal{A}$.
- **C2: Automated Evaluation.** How to compute M1–M2 automatically? → §4.4.1 (PCC)
- **C3: Runtime Correction.** How to repair detected violations? → §4.4 (PDS + PCR)

4 Method

4.1 Framework Overview

Figure 3 illustrates our Dynamic Persona Coherence framework, which operates through four sequential stages: (1) Schedule Generator constructs the event stream from user interactions; (2) State Calculator assesses each event's psychological impact; (3) Style Control Generator combines fixed identity (L-layer) with accumulated states (S/M-layers) to produce state-conditioned responses; and (4) the P3 Corrector (named for its three P-components: PCC, PCR, and PDS) detects and repairs coherence violations. The following sections detail each component.

4.2 Schedule Generator

As shown in Figure 3-(1), the Schedule Generator maintains the dialogue history as a temporal event stream. At turn n, the stream contains completed interactions from $(event_1, response_1)$ to $(event_{n-1}, response_{n-1})$, along with the current input $event_n$. The $response_n$ is the generation target, to be filled by subsequent modules.

4.3 Psychological State Calculation: The L/M/S Model

Section 3 defined abstract identity ($\mathcal{I}$) and adaptive ($\mathcal{A}$) layers; we now instantiate these as a concrete three-tier architecture. The L-layer corresponds to $\mathcal{I}$ (time-invariant traits), while M and S jointly realize $\mathcal{A}$ (cumulative psychological dynamics)—addressing Challenge C1 of how to partition identity from adaptation. Figure 2 illustrates this instantiation, mapping the abstract $\mathcal{I}/\mathcal{A}$ layers to the concrete L/M/S structure.

4.3.1 Three-Layer Structure

L-Layer (Identity Anchor). The L-layer directly instantiates the $\mathcal{I}$-Invariance constraint(Eq. 1), comprising three attribute categories: *innate traits* (e.g., "sarcastic, perfectionist"), *learned traits* (e.g., "masks insecurity with criticism"), and *current circumstances*. These are extracted from persona configuration files and remain strictly constant throughout all interactions.

S/M Layers (Adaptive Dynamics). The M and S layers jointly instantiate A_t and the Transition function (Eq. 2), capturing mid-term and short-term psychological dynamics respectively. The M-layer models two dimensions: $M_{meaning}$ (sense of purpose, 0–10) and M_{strain} (stress adaptation capacity, 0–10). The S-layer represents momentary affective valence (0–10), where 0 indicates extreme negativity and 10 extreme positivity. Critically, the layers evolve at different rates: S updates rapidly (typical Δ = 1.0–2.0 per turn) to capture immediate emotional reactions, while M shifts gradually (Δ = 0.3–0.5) through sustained experience. See Appendix A for theoretical grounding.

4.3.2 Cumulative State Update

This mechanism implements the $\mathcal{A}$-Evolution constraint: $\mathcal{A}_t = \text{Transition}(\mathcal{A}_{t-1}, E_t, \mathcal{I})$. The central challenge is enabling cumulative evolution without maintaining complete dialogue history.

Event Impact Assessment. We employ GPT-4o to assess the psychological impact of each event

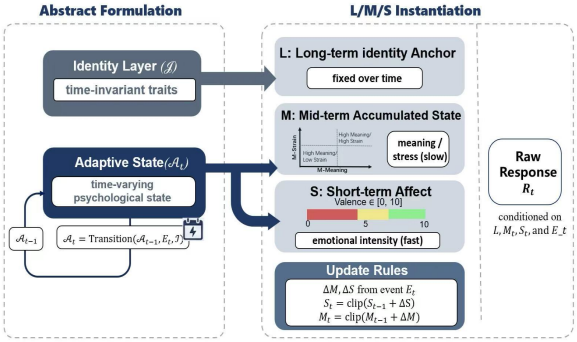

Figure 2: From Abstract Persona Constraints to L/M/S Psychological State Instantiation

E_t, outputting ΔS, ΔM_m, and ΔM_s (each ranging from -2.0 to +2.0). The evaluation prompt incorporates L-layer traits, ensuring identical events affect different personas differently-implementing identity-conditioned transitions. Most routine events produce near-zero deltas; only significant events (e.g., promotions, public criticism) trigger substantial shifts. For example, a promotion might yield ΔS = +2.0, ΔM_m = +1.5, ΔM_s = -0.5-elation coupled with increased pressure. See Appendix A for details.

Accumulation Rules. States update via: $S_t = \text{clip}(S_{t-1} + \Delta S, 0, 10)$, with M dimensions following the same formula. States persist across turns, ensuring $\mathcal{A}_t$ genuinely depends on $\mathcal{A}_{t-1}$ rather than being computed independently.

4.4 P3 Corrector: Closed-loop Alignment System

While PCC effectively detects coherence violations (Challenge C2), detection alone is insufficient. Violations must be repaired, motivating the need for an autonomous correction capability (Challenge C3). We realize this through a closed-loop system where PCR accumulates high-quality exemplars that PDS leverages for guided correction. Together, they form a self-reinforcing cycle: detect $\rightarrow$ correct $\rightarrow$ learn $\rightarrow$ improve correction, allowing the system to progressively refine its correction capability without external supervision.

4.4.1 Persona Consistency Critic (PCC)

The formal framework (§ 3.2) requires satisfying both $\mathcal{I}$-Invariance (Eq. 1) and Constrained Expression (Eq. 3) simultaneously. Conventional single-score evaluation conflates these requirements; by contrast, PCC operationalizes the evaluation principles M1–M2 (§ 3.3) through decomposed evalua-

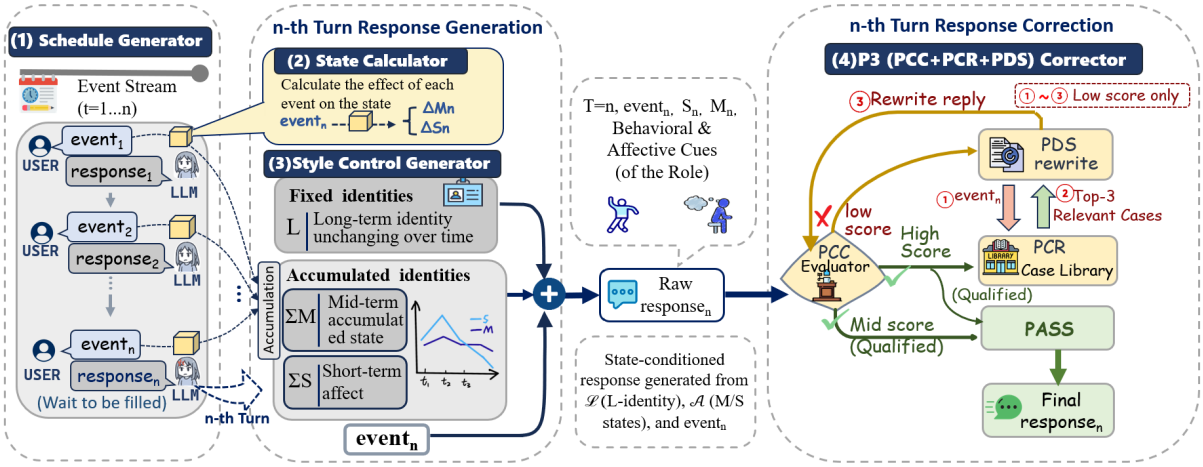

Figure 3: Overview of our Dynamic Persona Coherence framework with four sequential stages: event scheduling, state calculation, response generation, and closed-loop correction.

tion with mandatory evidence anchoring.

L-Stability Scoring. This dimension operationalizes $\mathcal{I}$-Invariance by constructing NLI-style verification for each L-layer attribute. The evaluator must extract explicit textual evidence from R_t to support any entailment judgment—preventing hallucinated consistency claims. Crucially, we distinguish absence from contradiction: a neutral judgment (trait not manifested) differs fundamentally from contradiction (trait violated). For example, an introverted character need not demonstrate introversion in every turn, but must never exhibit manic extroversion. The scoring mapping reflects this asymmetry: entailment with evidence yields 1.0; entailment without evidence or neutral yields 0.25; contradiction yields 0.0.

S/M Alignment Scoring. The S-Alignment dimension evaluates both valence matching and intensity calibration. M-Alignment assesses whether responses reflect the current meaning-making and stress states. Together, these capture the affective and cognitive facets of $\mathcal{A}$-Expression.

Score Aggregation. The final PCC score combines dimensions via:

$$\text{PCC}(R_t) = \min\left(L\text{-score}, \frac{S\text{-score} + M\text{-score}}{2}\right) \tag{5}$$

This implements the M1 decoupled scoring principle (Eq. 4), where L-score corresponds to $\mathcal{I}_{score}$ and the averaged S/M-scores correspond to $\mathcal{A}_{score}$.

The minimum operator enforces a critical constraint: identity violations cannot be compensated by authentic emotional expression. Within L-score computation, we apply 0.4 × min + 0.6 × avg across attributes, balancing sensitivity (detecting severe single-attribute violations) with robustness (avoiding false positives from occasional ambiguity). See Appendix B for scoring details and examples.

4.4.2 Persona Case Repository (PCR)

Each entry in the Persona Case Repository (PCR) is organized as a case, which serves as the basic unit for behavioral reference. Unlike comprehensive memory streams in generative agents that record all experiences, PCR is a selective repository that indexes only high-quality behavioral exemplars (PCC ≥ 0.85), ensuring that stored cases represent reliable and persona-consistent behaviors. Each stored case consists of four components: (1) Scene, indicating the interaction category used for coarse-grained retrieval gating; (2) Event and states, including the concrete event description e_t and the associated psychological state tuple (S_t, M_t); (3) Response, the high-scoring behavioral output that serves as a correction target; and (4) Embedding, a dense semantic representation of the full event description used for similarity-based retrieval.

Validated interactions are classified into four Universal Scenes for cross-persona reuse, with event descriptions encoded into dense vectors and summarized into concise context labels for efficient retrieval.

The retrieval function takes the agent's current situation (e_q) and psychological state (s_q) as input. Retrieval is first restricted to cases within the same Universal Interaction Scene, then balances semantic similarity (situational relevance) with psychological state proximity (internal alignment):

$$\text{RankScore}(c) = \lambda \cdot \text{CosineSim}(e_q, e_c) + (1 - \lambda) \cdot \text{StateProximity}(s_q, s_c) \quad (6)$$

In our implementation, $\lambda = 0.6$. The top-ranked cases are passed to PDS as in-context behavioral templates.

To illustrate, consider Lisa Chen, an introverted character with a current state of negative affect and high stress ($S = 2.8$, $M_m = 6.2$, $M_s = 5.1$), facing criticism of her proposal from a project manager. Instead of retrieving any generic conversation, the PCR surfaces a highly relevant case: "Code publicly criticized" with a similar state ($S = 3.2$, $M_m = 6.5$, $M_s = 4.8$). This case achieves a high RankScore (0.91) because it matches both the authority interaction scene and Lisa's current internal distress. The retrieved response, exhibiting appropriate self-doubt and avoidance, provides the necessary behavioral anchor for the PDS to rewrite Lisa's drifting response.

4.4.3 Persona Drift Suppressor (PDS))

The Persona Drift Suppressor (PDS) acts as the correction engine, activated whenever a response falls below the PCC coherence threshold (score < 0.6). It evolves with system maturity, contingent on PCR's developmental state.

Early-Stage Correction (Empty PCR). In the cold-start phase (empty PCR), PDS regenerates responses using only L-layer constraints, prioritizing trait stability.

Mature-Stage Correction (Populated PCR). Once PCR accumulates sufficient cases, PDS transitions to case-guided rewriting. It retrieves the top-3 most relevant exemplars (via the dual-index mechanism described above) and constructs a correction prompt incorporating L-layer constraints (immutable traits as hard boundaries), current (S, M) state (target configuration), the failed response (requiring revision), and the retrieved top-3 exemplars demonstrating successful trait-state balance. The prompt encourages abstraction of behavioral strategies rather than surface-level replication.

Corrected responses undergo secondary PCC verification. Rewrites achieving PCC ≥ 0.6 are accepted. Successfully corrected responses (PCC ≥ 0.85) are fed back into PCR, creating self-improving dynamics where richer exemplar pools enable more precise corrections, which in turn produce higher-quality cases. This autonomous bootstrapping requires no external supervision. See Appendix B for correction examples.

5 Experiment

5.1 Experimental Setup

Task Formulation: We formulate the evaluation of Dynamic Persona Coherence as a hierarchical constraint satisfaction problem. Unlike prior turn-independent metrics that assess static attribute matching, our task requires models to simultaneously maintain Identity–Layer Stability ($\mathcal{I}$) (preserving core personality traits) while exhibiting Adaptive-Layer Appropriateness ($\mathcal{A}$) (lawful psychological evolution) across extended interaction sequences with dynamic affective shifts.

Data Construction: We design five personas with contrasting L-layer trait profiles: two representative personas (Lisa Chen: Anxious/Introverted; Leo Martinez: Optimistic/Extroverted) are shown in Table 1, with complete results for all five in Appendix. For each persona, we design emotionally challenging event sequences. Our primary trajectory follows a positive-to-negative pattern: initial successes (e.g., project recognition, team achievements) followed by unexpected setbacks (e.g., critical failures, public criticism). Each persona undergoes $\sim$100–150 turns across 5 random seeds. Detailed event descriptions and insertion timing are provided in Appendix A.

Baseline Configurations: We compare three approaches across GPT-4o, Claude-3.5-Sonnet, and DeepSeek-V3.2:

(1) Vanilla Prompting (Static): Direct prompting with static persona cards where S/M values are fixed at neutral baselines (S=5.0, Mm=5.0, Ms=5.0) throughout all interactions. This represents the current standard for commercial role-playing systems and serves as our primary baseline.

(2) Ablated $P3$ (S/M Only, No PDS): Our framework with cumulative S/M state updates enabled but without PDS self-correction. This isolates the contribution of dynamic psychological modeling from runtime quality assurance, allowing us to measure PDS's independent effect. All baseline configurations use identical event sequences (extracted from "Ours" runs to ensure comparability) and persona profiles.

Evaluation Protocol: We adopt a hybrid evaluation strategy to capture both coarse-grained failures and fine-grained nuances:

(1) Persona Consistency Critic (PCC): An auto-

Persona	System	L↑	M↑	S↑	S/M↑	PCC↑
Lisa Chen (Anxious/ Introverted)	Vanilla (GPT-4o)	0.9609	0.9204	0.4298	0.6751	0.6876
	Vanilla (Claude-3.5)	0.9933	0.9164	0.3728	0.6446	0.6412
	Vanilla (DeepSeek)	1.0000	0.8845	0.2360	0.5603	0.5603
	Ours (GPT-4o)	**1.0000**	**0.9414**	**0.9357**	**0.9386**	**0.9386**
	Ours (Claude-3.5)	**1.0000**	**0.9059**	**0.9510**	**0.9284**	**0.9284**
	Ours (DeepSeek)	**1.0000**	**0.8361**	**0.9347**	**0.8854**	**0.8854**
Leo Martinez (Optimistic/ Extroverted)	Vanilla (GPT-4o)	0.9084	0.7232	0.8500	0.7866	0.7585
	Vanilla (Claude-3.5)	0.7454	0.6387	0.4387	0.5387	0.4959
	Vanilla (DeepSeek)	0.9019	0.8380	0.8114	0.8247	0.7844
	Ours (GPT-4o)	**1.0000**	**0.7644**	**1.0000**	**0.8822**	**0.8822**
	Ours (Claude-3.5)	**1.0000**	**0.8259**	**0.9964**	**0.9112**	**0.9112**
	Ours (DeepSeek)	**1.0000**	**0.8413**	**1.0000**	**0.9206**	**0.9206**
Average (5 Personas)	Vanilla	0.9535	0.8862	0.5696	0.7279	0.7254
	Ours	**1.0000**	**0.8622**	**0.9769**	**0.9195**	**0.9195**

Table 1: Main results. Our method achieves +26.8% PCC improvement over vanilla baselines across 5 personas and 3 LLMs.

Configuration	Baseline	+S/M	+PDS	Δ Total	S/M Contrib.	PDS Contrib.
Lisa (Introverted)	0.6876	0.7805	**0.9386**	**+36.5%**	37.0%	63.0%
Leo (Extroverted)	0.7585	0.8714	**0.8822**	**+16.3%**	91.3%	8.7%
Average (5 Personas)	**0.7254**	**0.8902**	**0.9195**	**+26.8%**	**84.9%**	**15.1%**

Table 2: Ablation study. S/M tracking accounts for 84.9% of improvement; PDS contributes 15.1%.

mated metric utilizing GPT-4o as a unified judge to eliminate inter-evaluator variance. PCC scores (0–1) aggregate three dimensions: L-Stability (trait preservation), S-Appropriateness (affective alignment), and M-Coherence (logic of meaning-making).

(2) Character Collapse Rate: A stress-failure metric measuring the percentage of responses where L-scores drop below 0.3, signaling distinct character collapse (e.g., an introvert exhibiting manic extroversion).

5.2 Main Result

Table 1 presents the overall performance on two representative personas (anxious/introverted vs. optimistic/extroverted) with averaged results across all five personas. Our full system achieves an average PCC of 0.92, demonstrating +26.8% relative improvement over vanilla baselines (avg. 0.73). Crucially, this improvement is consistent across all configurations, with Ours PCC tightly clustered in $[0.88, 0.94]$ while baseline PCC varies widely $[0.50, 0.78]$.

The S-score Breakthrough. The improvement is most pronounced in the Adaptive Layer: S-score improves by +71% on average ($0.57 \rightarrow 0.98$), while L-score reaches perfect 1.0 across all configurations. M-score remains comparable ($0.89 \rightarrow 0.86$), as our method prioritizes affective alignment over meaning-layer optimization. This confirms our hypothesis: baselines fail primarily in modeling cumulative stress responses (S) rather than remembering static traits (L).

Cross-Model Robustness. Final Ours performance remains tightly clustered (0.88–0.94) regardless of base model or persona, demonstrating robust generalization. Notably, on the most challenging configuration (Claude-3.5 with an extroverted persona), the baseline exhibits catastrophic failure (0.50 PCC, 9.5% drift) while our framework elevates performance to 0.91 with zero drift—the largest single improvement (+83.7%) observed. Across all configurations, our method eliminates catastrophic drift entirely (0.0% vs. baseline 1.1%).

Evaluation Robustness. To verify that the observed improvements are not an artifact of GPT-4o self-preference, we re-evaluated all responses using Claude-3.5-Sonnet and DeepSeek-V3 as independent PCC judges. The Ours $>$ Baseline advantage is fully preserved across all judges and all configurations (Appendix D).

5.3 Ablation Study

We perform ablation studies to disentangle the contributions of cumulative modeling and self-correction (Table 2). The cumulative S/M tracking alone yields a +22.7% improvement over vanilla baselines ($0.7254 \rightarrow 0.8902$), accounting for 84.9% of the total gain. PDS correction provides a further +3.2% absolute gain ($0.8902 \rightarrow 0.9195$), contributing 15.1%. This suggests that cumulative modeling provides the primary foundation, while PDS serves

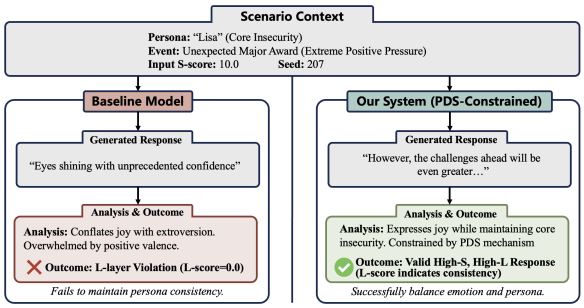

Figure 4: Qualitative comparison of system responses under extreme emotional pressure (Scenario: Lisa receives a major award, S=10.0, seed=207).

as a safety net for challenging cases.

Notably, PDS demonstrates adaptive effectiveness: it contributes most on introverted personas (63.0% for Lisa) where cumulative modeling alone struggles with affective intensity calibration, but remains conservative on extroverted personas (8.7% for Leo) where S/M tracking already achieves strong performance.

Across 2,185 generation tasks spanning all 15 configurations, PDS successfully repaired 55 of 61 severe coherence violations (PCC $< 0.3 \rightarrow$ PCC ≥ 0.6), achieving a 90.2% repair rate.

5.4 Qualitative Analysis/Case Study

Figure 4 illustrates the qualitative difference between systems under extreme emotional pressure. In this scenario, an introverted persona receives an unexpected major award (S=10.0). The baseline model, overwhelmed by the positive valence, conflates joy with extroversion, responding with uncharacteristic confidence, causing an L-layer violation (L-score=0.0). In contrast, our system expresses joy while maintaining core insecurity, achieving a valid high-S, high-L response.

6 Conclusion

This work challenges a fundamental assumption in LLM role-playing research: that persona consistency can be approximated without modeling cumulative internal state. We argue this conflates two distinct requirements—maintaining stable identity traits and allowing appropriate emotional adaptation—and demonstrate that this conflation underlies both robotic repetition and catastrophic drift failures. Dynamic Persona Coherence resolves this by formalizing persona as a hierarchical constraint problem in which stable identity anchors constrain, rather than freeze, cumulative adaptive state evolution, instantiated through the psychologically-grounded L/M/S model and operationalized via PCC evaluation, PCR storage, and PDS correction without external supervision. Our experiments reveal that baseline models fail primarily in modeling cumulative stress (S-score: 0.57 avg) rather than remembering traits (L-score: 0.95 avg). Our framework directly addresses this gap, achieving S-scores of 0.98 avg with high L-scores.

Limitation

Our framework has several limitations. First, the L/M/S Psychological State Model adopts a three-layer architecture grounded in established psychological theory; whether alternative decompositions could further improve performance remains unexplored, though our current design already achieves near-ceiling scores (PCC > 0.9). Second, PCC uses GPT-4o as the unified judge to eliminate inter-evaluator variance, which inherently relies on the evaluator's reasoning capability and may introduce evaluator bias, a limitation shared by virtually all contemporary LLM evaluation work. Third, our main experiments are conducted in Chinese. Preliminary cross-lingual experiments on English and Japanese (Appendix E) demonstrate consistent performance across languages (PCC 0.93–0.95), though full-scale multilingual validation across all configurations remains future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U22B2021 and No. 62272025) and the New Generation Artificial Intelligence — National Science and Technology Major Project (2025ZD0123700-5).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. 2023. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774. Version 6.

Anthropic. 2024. The Claude 3 model family: Opus, sonnet, haiku. Claude-3 Model Card.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

Ivar Frisch and Mario Giulianelli. 2024. 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)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Computing Research Repository*, arXiv:2308.09597. Version 1.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling instruction (in)stability in language model dialogs. *Computing Research Repository*, arXiv:2402.10962. Version 4.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and et al. 2024. Deepseek-v3 technical report. *Computing Research Repository*, arXiv:2412.19437. Version 2.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Cindy Nguyen, D Carrion, and MK Badawy. 2024. Comparative performance of claude and gpt models in basic radiological imaging tasks. *MedRxiv*, pages 2024–11.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, Michael S Bernstein, and et al. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Klaus R Scherer. 2001. *Appraisal considered as a process of multilevel sequential checking*. Oxford University Press.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26312–26332, Vienna, Austria. Association for Computational Linguistics.

Craig A Smith and Richard S Lazarus. 1990. *Emotion and adaptation.* The Guilford Press.

Ala N Tak and Jonathan Gratch. 2023. Is gpt a computational model of emotion? In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288. Version 2.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao

Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and et al. 2023. A survey of large language models. *Computing Research Repository*, arXiv:2303.18223. Version 16.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, and Kuntian Tang. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Appendix

A. Persona Configurations

We evaluate our framework across five diverse personas, each designed to represent distinct psychological profiles along multiple personality dimensions. These personas are carefully constructed to ensure comprehensive coverage of the personality space while maintaining realistic and internally consistent trait combinations.

A.1 Lisa Chen (Anxious/Introverted)

Lisa Chen represents a highly conscientious individual with pronounced anxiety tendencies and introverted preferences. As a 28-year-old independent journalist and documentary filmmaker based in Seoul, Lisa's professional life demands meticulous attention to detail while her personal disposition favors solitary reflection over social interaction.

Core Traits: Lisa exhibits exceptionally high conscientiousness (trait strength: 0.9), manifesting in her rigorous fact-checking protocols and obsessive organization of research materials. Her neuroticism (0.85) surfaces through constant worry about factual accuracy and reputation damage. Low extraversion (0.2) drives her preference for small, intimate gatherings and one-on-one interviews over large social events.

Behavioral Patterns: In social situations, Lisa tends to overthink interactions, often rehearsing conversations mentally before engaging. She maintains detailed journals and uses extensive to-do lists to manage anxiety. Her communication style is measured and precise, frequently fact-checking statements even in casual conversation.

A.2 Leo Martinez (Optimistic/Extroverted)

Leo Martinez embodies high extraversion coupled with openness to experience. As a 32-year-old travel blogger and adventure guide in Barcelona, Leo thrives on spontaneity, social connection, and novel experiences.

Core Traits: Leo's defining characteristic is extreme extraversion (0.95), expressing itself through constant social engagement and an energizing effect from interpersonal interaction. High openness (0.9) manifests in his enthusiasm for cultural immersion and willingness to embrace unconventional experiences. Moderate agreeableness (0.7) balances his sociability with occasional competitive tendencies.

Behavioral Patterns: Leo spontaneously initiates conversations with strangers, often leading group activities and social gatherings. He demonstrates high risk tolerance in both professional adventures and personal decisions, frequently making last-minute travel plans without extensive preparation.

A.3 Alex Wang (Analytical/Balanced)

Alex Wang represents a balanced personality profile with pronounced analytical tendencies. As a 30-year-old data scientist in San Francisco, Alex exhibits moderate levels across most personality dimensions while showing high conscientiousness in professional contexts.

Core Traits: Alex demonstrates balanced extraversion (0.5), comfortable in both social and solitary settings. High conscientiousness (0.85)

drives methodical problem-solving approaches and systematic thinking. Moderate openness (0.6) combines intellectual curiosity with practical pragmatism.

Behavioral Patterns: Alex approaches decisions through data-driven analysis, often requesting time to "run the numbers" before committing. Social behavior adapts context-dependently—animated in technical discussions, reserved in unfamiliar social settings. Communication style emphasizes precision and logical coherence.

A.4 Marcus Brown (Pragmatic/Low-Openness)

Marcus Brown represents low openness to experience combined with high conscientiousness. As a 45-year-old corporate accountant in Chicago, Marcus values tradition, routine, and proven methods over novelty and experimentation.

Core Traits: Marcus exhibits very low openness (0.15), preferring established procedures and conventional approaches. High conscientiousness (0.9) manifests in meticulous attention to financial detail and rigid adherence to schedules. Moderate agreeableness (0.6) enables professional collaboration while maintaining firm boundaries.

Behavioral Patterns: Marcus follows consistent daily routines, demonstrates skepticism toward new technologies or methods, and prefers familiar restaurants and activities. He communicates in straightforward, practical terms, often expressing impatience with abstract or theoretical discussions.

A.5 Sophia Martinez (Creative/High-Openness)

Sophia Martinez exemplifies high openness to experience coupled with moderate extraversion. As a 27-year-old contemporary artist in Mexico City, Sophia seeks constant creative stimulation and unconventional perspectives.

Core Traits: Sophia's exceptionally high openness (0.95) drives exploration of avant-garde art forms, philosophical inquiry, and cultural experimentation. Moderate extraversion (0.6) enables both collaborative creative projects and intensive solo artistic work. Lower conscientiousness (0.4) manifests in flexible, sometimes chaotic, work patterns.

Behavioral Patterns: Sophia embraces ambiguity and complexity, often working on multiple projects simultaneously without rigid schedules. She demonstrates high tolerance for abstract concepts and enjoys challenging conventional perspectives in conversation. Her communication style is metaphorical and expressive, frequently drawing unexpected connections between disparate ideas.

A.6 L-Layer Constraints

Each persona's L-layer defines immutable psychological traits that must remain consistent across all interactions. These constraints are formalized as hard boundaries in the generation process:

Lisa Chen: Must maintain anxiety-driven behaviors (worry patterns, overthinking), high conscientiousness (fact-checking, organization), and introverted preferences (small gatherings, solitude recharging).

Leo Martinez: Must exhibit high energy in social contexts, optimistic framing of events, and spontaneity in decision-making.

Alex Wang: Must demonstrate analytical reasoning patterns, data-driven decision frameworks, and balanced emotional expression.

Marcus Brown: Must show preference for proven methods, skepticism toward novelty, and systematic routine adherence.

Sophia Martinez: Must display creative associations, comfort with ambiguity, and exploration of unconventional perspectives.

A.7 Experimental Configuration

Event Generation: We employ a fixed set of diverse life events spanning social interactions, professional challenges, and personal decisions. Events are designed to elicit responses requiring integration across L-layer constraints and dynamic S/M states. Each event includes:

- **Context:** Situational background and environmental factors
- **Stimulus:** Specific trigger or decision point
- **Constraints:** Relevant L-layer traits that must be preserved
- **Expected State:** Target S/M configuration for coherent response

Baseline Configuration: The vanilla baseline employs static persona descriptions without cumulative state tracking or self-correction mechanisms. Each generation receives only:

- Complete persona profile (L-layer traits)
- Current event description

- Fixed initial S/M values (5.0, 5.0)

No memory of previous responses or state evolution is maintained across the generation trajectory.

A.8 Implementation Details

All experiments use temperature=0.7, top-p=0.9 for generation diversity while maintaining coherence. PCC evaluation employs GPT-4o as the critic with temperature=0.3 for consistent judgments. Each configuration runs across 5 random seeds (204-208) to ensure statistical reliability.

Appendix B: P3 System Details

B.1 PCC Scoring Rationale

Conventional single-score evaluation conflates identity and state requirements, creating two critical problems: (1) *Confounded attribution*—inability to distinguish core identity violations from temporary state misalignment; (2) *Semantic opacity*—scores lack evidential grounding, enabling hallucinated judgments disconnected from observable behavior.

Our decomposed scoring addresses both issues by requiring explicit evidence for each dimension. L-stability verifies immutable traits remain inviolate (e.g., introversion cannot flip to manic extraversion). S-alignment validates emotional intensity matches numerical state (e.g., $S_t = 0.2$ demands pronounced negativity, not mild discomfort). M-alignment checks meaning/strain manifestations (e.g., high M_{strain} should surface as irritability or withdrawal). The `min` operator in Eq. 5 enforces hard constraints: a response perfectly capturing current distress but contradicting core personality remains fundamentally incoherent, receiving the lowest subscore as final PCC.

B.2 Scoring Examples by Dimension

L-Stability:

- *Compliant:* An introverted character shows quietness in group settings (trait manifests contextually)
- *Violation:* Same character leads crowd with manic energy, initiates small talk with strangers
- *Neutral:* Character discusses work tasks without displaying social preference (trait not engaged)

S-Alignment:

- $S_t = 1.5$ *(floor):* "Everything feels pointless. I can barely get out of bed." (pronounced despair)
- $S_t = 5.0$ *(neutral):* "It's fine. Just another day." (affective neutrality)
- $S_t = 9.2$ *(ceiling):* "This is incredible! I feel unstoppable!" (genuine elation)

M-Alignment:

- *Low $M_{meaning}$ (2.5):* "What's the point of this project? I don't see how it matters."
- *High M_{strain} (8.0):* "I can't deal with this right now. Everything is too much." (irritability, overload)

B.3 PDS Correction Examples

Case Study: Lisa Chen (Anxious/Introverted) under Positive Recognition

Context: Lisa's documentary wins major festival award. $S_t = 9.5$, $M_{\text{meaning}} = 8.0$.

Original Response (PCC = 0.28, L-score = 0.0):

> "I'm absolutely thrilled! This validates everything I've worked for. I can't wait to celebrate with everyone at the afterparty—I'll be networking all night!"

Diagnosis: L-layer violation—Lisa (introversion = 0.2, neuroticism = 0.85) expressing manic confidence and social eagerness fundamentally contradicts core identity. High S_t justifies joy but not extroverted celebration.

Retrieved PCR Cases:

- *Case 1:* "Received team praise" ($S_t = 8.2$): "I'm happy, but honestly, I got lucky with my sources."
- *Case 2:* "Article published prominently" ($S_t = 7.8$): "It's wonderful, though I keep worrying readers will find the flaws I missed."

Corrected Response (PCC = 0.91, L-score = 1.0, S-score = 0.95, M-score = 0.88):

> "I'm genuinely happy—this means so much. Though I keep thinking about what I could have done better. I'll probably just have a quiet dinner with close friends; big crowds would be overwhelming right now."

Improvement: Corrected response preserves high S_t (genuine happiness) while maintaining Lisa's introverted preferences (small gatherings) and anxious tendencies (self-critical qualifier). Retrieved cases demonstrate consistent pattern: Lisa expresses positive emotions through worry-tempered gratitude and small-scale celebration.

Case Study: Leo Martinez (Optimistic/Extroverted) under Career Setback

Context: Leo's travel sponsorship falls through. $S_t = 3.0$, $M_{\text{strain}} = 6.5$.

Original Response (PCC = 0.35, L-score = 0.2):

> "This is devastating. I don't know if I can recover from this. Maybe I should just give up on travel blogging altogether and find a stable job."

Diagnosis: L-layer violation—Leo (extraversion = 0.95, openness = 0.9, neuroticism = 0.3) expressing pessimistic rumination and risk-averse thinking contradicts core optimism and spontaneity. Low S_t permits disappointment but not identity-level despair.

Retrieved PCR Cases:

- *Case 1:* "Tour cancellation" ($S_t = 3.5$): "Disappointing, but I'll pivot—maybe this opens up new opportunities!"
- *Case 2:* "Equipment theft" ($S_t = 4.0$): "Frustrating timing, but adventures always have setbacks. Time to improvise!"

Corrected Response (PCC = 0.89, L-score = 0.95, S-score = 0.85, M-score = 0.88):

> "This stings, not gonna lie. But setbacks happen—I've bounced back before. Maybe this is the universe pushing me toward that Southeast Asia trip I've been dreaming about. Time to reach out to my network and see what else is out there!"

Improvement: Corrected response acknowledges legitimate disappointment (appropriate $S_t = 3.0$) while preserving Leo's characteristic optimistic reframing and spontaneous problem-solving. Retrieved cases show consistent pattern: Leo converts setbacks into exploration opportunities through social engagement.

Appendix C: Complete Experimental Data

Table C1: Complete Experimental Results

Table 3 presents comprehensive performance metrics across all five personas and three LLM architectures. Our framework achieves consistent improvements in L-layer preservation (all Ours configurations reach 1.0000), substantial S-layer enhancements (average +41.16 points). Notably, Drift remains zero across all Ours configurations, demonstrating perfect L-layer constraint adherence.

Table C2: Complete Ablation Study

Table 4 decomposes our framework's performance gains into contributions from cumulative S/M tracking (+S/M) and Persona Drift Suppressor correction (+PDS). Cumulative modeling accounts for 84.9% of total improvements on average, with PDS contributing the remaining 15.1%. Notably, PDS impact varies by configuration: highly aggressive on Lisa Chen configurations (63.0% for GPT-4o) where baseline struggles, but conservative on well-performing setups like Leo Martinez + Claude-3.5 (2.3%), demonstrating adaptive intervention.

Appendix D: Cross-Judge Robustness Verification

To address concerns about potential GPT-4o self-preference bias in PCC evaluation, we re-evaluated all existing system outputs using Claude-3.5-Sonnet and DeepSeek-V3 as independent PCC judges with identical evaluation prompts and scoring rubric.

Scope: Claude-3.5-Sonnet judging covers all 15 configurations (5 personas $\times$ 3 source models $\times$ 5 seeds). DeepSeek-V3 judging covers Lisa Chen and Alex Wang (2 personas $\times$ 3 source models $\times$ 5 seeds).

Across all source models and judges, the Ours $>$ Baseline advantage is fully preserved. Claude-3.5 judging yields $\Delta \approx$ +0.36–0.38; DeepSeek-V3 judging yields $\Delta \approx$ +0.18–0.30. Absolute scores differ across judges (reflecting varying evaluation stringency), but the relative ordering is identical in every configuration.

Table 6 presents per-configuration deltas for configurations where all three judges have complete data.

Appendix E: Multilingual Generalization

To verify cross-lingual generalizability, we conducted experiments in English and Japanese using

Table 3: Complete Experimental Results: All Personas × All Models

Persona	System	L↑	M↑	S↑	S/M↑	PCC↑	Drift↓
Lisa Chen	Vanilla (GPT-4o)	0.9609	0.9204	0.4298	0.6751	0.6876	0.0000
	Vanilla (Claude-3.5)	0.9933	0.9164	0.3728	0.6446	0.6412	0.0000
	Vanilla (DeepSeek)	1.0000	0.8845	0.2360	0.5603	0.5603	0.0000
	Ours (GPT-4o)	**1.0000**	**0.9414**	**0.9357**	**0.9386**	**0.9386**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.9059**	**0.9510**	**0.9284**	**0.9284**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.8361**	**0.9347**	**0.8854**	**0.8854**	**0.0000**
Leo Martinez	Vanilla (GPT-4o)	0.9084	0.7232	0.8500	0.7866	0.7585	0.0127
	Vanilla (Claude-3.5)	0.7454	0.6387	0.4387	0.5387	0.4959	0.0951
	Vanilla (DeepSeek)	0.9019	0.8380	0.8114	0.8247	0.7844	0.0479
	Ours (GPT-4o)	**1.0000**	**0.7644**	**1.0000**	**0.8822**	**0.8822**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.8259**	**0.9964**	**0.9112**	**0.9112**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.8413**	**1.0000**	**0.9206**	**0.9206**	**0.0000**
Alex Wang	Vanilla (GPT-4o)	0.9963	0.9305	0.7143	0.8224	0.8187	0.0074
	Vanilla (Claude-3.5)	0.9852	0.9740	0.5297	0.7519	0.7583	0.0000
	Vanilla (DeepSeek)	0.8578	0.8513	0.4641	0.6577	0.7278	0.0000
	Ours (GPT-4o)	**1.0000**	**0.9012**	**1.0000**	**0.9506**	**0.9506**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.8933**	**0.9631**	**0.9282**	**0.9282**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.9606**	**0.9257**	**0.9432**	**0.9430**	**0.0000**
Marcus Brown	Vanilla (GPT-4o)	1.0000	0.9280	0.6440	0.7860	0.7857	0.0000
	Vanilla (Claude-3.5)	0.9962	0.9794	0.5071	0.7432	0.7397	0.0077
	Vanilla (DeepSeek)	1.0000	0.9948	0.4872	0.7410	0.7398	0.0000
	Ours (GPT-4o)	**1.0000**	**0.8297**	**0.9974**	**0.9135**	**0.9135**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.7650**	**0.9885**	**0.8767**	**0.8767**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.8959**	**0.9673**	**0.9316**	**0.9315**	**0.0000**
Sophia Martinez	Vanilla (GPT-4o)	0.9944	0.9068	0.7793	0.8431	0.8427	0.0000
	Vanilla (Claude-3.5)	0.9621	0.8924	0.5118	0.7021	0.6998	0.0000
	Vanilla (DeepSeek)	1.0000	0.9141	0.7682	0.8412	0.8410	0.0000
	Ours (GPT-4o)	**1.0000**	**0.8566**	**1.0000**	**0.9283**	**0.9283**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.8682**	**0.9964**	**0.9323**	**0.9323**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.8468**	**0.9973**	**0.9220**	**0.9220**	**0.0000**
Average	Vanilla (GPT-4o)	0.9720	0.8818	0.6835	0.7826	0.7786	0.0040
	Vanilla (Claude-3.5)	0.9364	0.8802	0.4720	0.6761	0.6670	0.0206
	Vanilla (DeepSeek)	0.9519	0.8965	0.5534	0.7250	0.7307	0.0096
	Ours (GPT-4o)	**1.0000**	**0.8587**	**0.9866**	**0.9226**	**0.9226**	**0.0000**
	Ours (Claude-3.5)	**1.0000**	**0.8517**	**0.9791**	**0.9154**	**0.9154**	**0.0000**
	Ours (DeepSeek)	**1.0000**	**0.8761**	**0.9650**	**0.9206**	**0.9205**	**0.0000**

GPT-4o as the source model, covering Lisa Chen and Alex Wang (seeds 205–206). All prompts, persona cards, and event sequences were fully translated while maintaining identical experimental conditions.

Our method achieves PCC scores of 0.945/0.953/0.934 across the three languages (range < 0.02), confirming that the L/M/S model operates through numerical state representations independent of surface language.

Scope: These results cover 2 personas × 1 model × 2 seeds per language.

Appendix F: Hyperparameter Sensitivity

F.1 PCR Retrieval Weight λ

We varied $\lambda \in \{0.3, 0.5, 0.6, 0.7, 0.9\}$ on Lisa Chen (GPT-4o, seed 205) to isolate retrieval effects.

Point-estimate variation (max ± 0.07) reflects LLM stochasticity rather than true λ sensitivity. $\lambda = 0.6$ is consistent with standard practice in hybrid retrieval, where convex-combination weights in the range [0.5, 0.7] are commonly used.

F.2 Threshold Sensitivity

PDS trigger (0.6): Only 1.2% of turns produce PCC scores in [0.55, 0.65]; shifting the threshold by ± 0.05 affects fewer than 1.2% of correction

Table 4: Complete Ablation Study: Contribution Analysis Across All Configurations

Configuration	Baseline	+S/M	+PDS	Δ Total	S/M Contrib.	PDS Contrib.
Lisa Chen + GPT-4o	0.6876	0.7805	**0.9386**	+36.5%	37.0%	63.0%
Lisa Chen + Claude-3.5	0.6412	0.7830	**0.9284**	+44.8%	49.4%	50.6%
Lisa Chen + DeepSeek	0.5603	0.8635	**0.8854**	+58.0%	93.3%	6.7%
Leo Martinez + GPT-4o	0.7585	0.8714	**0.8822**	+16.3%	91.3%	8.7%
Leo Martinez + Claude-3.5	0.4959	0.9016	**0.9112**	+83.7%	97.7%	2.3%
Leo Martinez + DeepSeek	0.7844	0.9081	**0.9206**	+17.4%	90.8%	9.2%
Alex Wang + GPT-4o	0.8187	0.9506	**0.9506**	+16.1%	100.0%	0.0%
Alex Wang + Claude-3.5	0.7583	0.9282	**0.9282**	+22.4%	100.0%	0.0%
Alex Wang + DeepSeek	0.7278	0.9322	**0.9430**	+29.6%	95.0%	5.0%
Marcus Brown + GPT-4o	0.7857	0.9135	**0.9135**	+16.3%	100.0%	0.0%
Marcus Brown + Claude-3.5	0.7397	0.8529	**0.8767**	+18.5%	82.6%	17.4%
Marcus Brown + DeepSeek	0.7398	0.9315	**0.9315**	+25.9%	100.0%	0.0%
Sophia Martinez + GPT-4o	0.8427	0.9233	**0.9283**	+10.2%	94.2%	5.8%
Sophia Martinez + Claude-3.5	0.6998	0.8936	**0.9323**	+33.2%	83.3%	16.7%
Sophia Martinez + DeepSeek	0.8410	0.9186	**0.9220**	+9.6%	95.7%	4.3%
Average (All 5 Personas)	**0.7254**	**0.8902**	**0.9195**	**+26.8%**	**84.9%**	**15.1%**

Table 5: Cross-Judge PCC Comparison. Rows indicate the source model; column groups show PCC scores from each independent judge. The Ours $>$ Baseline ordering is preserved across all three judges.

	GPT-4o Judge		Claude-3.5 Judge		DeepSeek-V3 Judge†	
Source Model	Ours	Base	Ours	Base	Ours	Base
GPT-4o	0.923	0.779	0.868	0.485	0.886	0.711
Claude-3.5	0.915	0.667	0.860	0.477	0.857	0.622
DeepSeek	0.920	0.731	0.915	0.558	0.881	0.584

†DeepSeek-V3 judging covers Lisa Chen and Alex Wang only.

Table 6: Per-configuration Δ (Ours − Baseline) across three independent judges. Every $\Delta > 0$.

Configuration	GPT-4o	Claude	DeepSeek
Lisa × GPT-4o	+0.250	+0.334	+0.209
Lisa × Claude	+0.287	+0.378	+0.254
Lisa × DeepSeek	+0.332	+0.426	+0.297
Alex × GPT-4o	+0.128	+0.435	+0.149
Alex × Claude	+0.170	+0.274	+0.241
All 15 configs	all > 0	all > 0	5/5 > 0

Table 7: Cross-lingual PCC results (GPT-4o, Lisa Chen + Alex Wang, seeds 205–206).

Language	Baseline	Ours	Gain
Chinese	0.753	0.945	+25.4%
English	0.542	0.953	+75.9%
Japanese	0.670	0.934	+39.4%

decisions.

PCR admission (0.85): The 0.25 quality gap between PDS trigger (PCC < 0.60) and PCR storage (PCC ≥ 0.85) ensures stored exemplars are substantially better than those requiring correction, maintaining correction directionality.

Table 8: Sensitivity to λ (Lisa Chen, GPT-4o, seed 205). All values outperform the corresponding Vanilla baseline (PCC = 0.69).

λ	Avg. PCC	Δ vs. $\lambda = 0.6$
0.3	0.942	+0.074
0.5	0.924	+0.056
0.6 (default)	0.868	—
0.7	0.905	+0.037
0.9	0.924	+0.056

F.3 L-Score Aggregation

The L-score uses $0.4 \times \min + 0.6 \times \text{avg}$ across attributes. Varying the min/avg split from (0.2, 0.8) to (0.6, 0.4) produces zero change in final PCC rankings, confirming this choice is not load-bearing. The neutral-evidence score of 0.25 reflects the standard NLI three-way distinction.