

PersonaForge: Psychology-Grouped Dual-Process Architecture for Personality-Consistent Role-Playing Agents

Jizhou Tong^{†*}
Wuhan University
fqwqf6@whu.edu.cn

Sirui Zou[†]
Southwestern University of Finance and Economics
siruizou2005@gmail.com

Abstract

Large Language Models excel at role-playing but struggle to maintain consistent personalities across *extended multi-turn interactions*. We propose **PersonaForge**, combining (1) a **three-layer personality architecture** grounded in psychological theory and (2) a **dual-process generation mechanism** inspired by cognitive science. We test two falsifiable claims: **Claim 1 (Orthogonality)**: Psychology-grounded dimensions (Big Five + Defense Mechanisms) provide more orthogonal constraints than natural language descriptions, reducing long-dialogue drift. **Claim 2 (Integration Necessity)**: High-dimensional personality constraints create “production interference” requiring a cognitive workspace (Inner Monologue) to resolve—removing it degrades performance *below* simpler baselines. Experiments on **88 characters** demonstrate: (1) +19.4% personality consistency (PC) over the Structured-CoT baseline, with human correlation $r = 0.82$, (2) reduced drift over 50-turn conversations (6.3% vs. 24.8% baseline), and (3) +64.7% defense mechanism manifestation. External validation on **RoleBench** confirms generalization (73.2% win-rate, drift 8.4% vs. 20.4%). Selective dual-process activation achieves 96% of full-system performance with only **13.4% token overhead**. Human evaluation confirms more authentic and psychologically coherent character behaviors. Code and data: <https://github.com/fqwQf/PersonaForge>.

1 Introduction

Role-playing with Large Language Models (LLMs) has emerged as a compelling application, enabling immersive storytelling and virtual companionship (Chen et al., 2024). However, maintaining consistent character personalities across extended multi-turn interactions remains a

fundamental challenge. Current approaches often suffer from:

1. **Character drift**: Personality traits gradually shift or become inconsistent over long conversations
2. **Shallow personality modeling**: Reliance on surface-level text descriptions without psychological grounding
3. **Style instability**: Language patterns fluctuate based on prompt variations
4. **Static representations**: Inability to model dynamic emotional states

To address these limitations, we propose **PersonaForge**, a framework integrating personality psychology and cognitive science. Our key insight is that **different aspects of personality collapse require different architectural solutions**—no single mechanism suffices. To address internal conflicts to achieve psychological integration, our architecture explicitly models both conscious expression and the suppressed cognitive processes beneath. To our knowledge, this is **among the first computational frameworks** to operationalize Vaillant’s defense mechanisms within LLM agents as programmable cognitive strategies.

SFT Enhancer and Inference-Time Guardrail. While SFT achieves stylistic mimicry (Shao et al., 2023), **low-resource SFT** (typical in cold-start scenarios) often captures the “surface voice” without the “cognitive core,” leading to rigidity in long contexts. PersonaForge can serve as an **Inference-time Guardrail** or **SFT Enhancer**: even effectively fine-tuned models benefit from our dual-process architecture to maintain consistency under stress. Furthermore, our approach solves the **cold-start problem** through **Automated Parameter Acquisition**: users need only provide raw text (e.g., a Wiki link), from which our pipeline extracts the Big Five and Defense Mechanism parameters, enabling high-fidelity role-play without

[†]Equal contribution. *Corresponding author.

Case Study: Psychological Defense Mechanism (Tyrion Lannister)

Context: Tywin Lannister calls Tyrion a "disgrace" to the family.

SFT Baseline (template-locked):

"I suppose I should be offended, but then again, I've always been a bit of a disappointment..."

Analysis: Acknowledges insult but lacks the characteristic wit. **Psychologically Flat.**

Ours (With Humor Defense):

Inner Monologue: [Defense: Humor → "Imagine him in a gilded bathtub..."]

Response: "Imagine him, in a gilded bathtub filled with rose petals, sipping champagne and humming a jaunty tune—quite the spectacle, wouldn't you say? Shall we raise a glass to my father?"

Analysis: Deflects the ego-threat by reducing the aggressor to absurdity. **Psychologically Authentic.**

Figure 1: Impact of Defense Mechanisms. The architecture enables characters to *face their true selves*, transforming generic emotional reactions into character-specific coping strategies.

training.

Our contributions:

Two Testable Research Claims. We frame our design around two falsifiable claims: **(1) The Orthogonality Hypothesis**—psychology-grounded dimensions (Big Five + Defense Mechanisms) provide more orthogonal constraints than natural language descriptions. Evidence: "Generic Structured" ablation (Table 3) shows $\Delta PC = -0.18$ when replacing Big Five/DM with natural language. **(2) The Integration Necessity Hypothesis**—high constraint density causes "production interference" without a cognitive workspace. Evidence: "w/o Dual-Process" achieves PC 0.57 vs. Structured-CoT's 0.72, confirming the three-layer profile *requires* the Inner Monologue.

Why Defense Mechanisms Are Not "Situation Templates." DMs are **cognitive distortion procedures**, not behavioral scripts—they specify *how to perceive* stressors, not just *what to say*. The same DM produces radically different outputs depending on stressor content. Ablation shows DM removal degrades *activation precision* from 92% to 61%. We model DMs as *pre-conscious cognitive routines* (Vaillant, 1994)—the Inner Monologue represents what surfaces to working memory, not the unconscious process itself. Following Marr's levels, we implement an *algorithmic* approximation of a *computational* theory of per-

sonality.

Three-Layer Personality Architecture. We introduce a **functionally decomposed** architecture: **Core Traits** (Big Five + Defense Mechanisms) prevent *trait drift*; **Speaking Style Matrix** prevents *style collapse*; **Dynamic State** prevents *context blindness* by tracking emotional/relational evolution. Inspired by Kahneman's dual-system theory (Kahneman, 2011), we implement *Think-then-Speak* that activates only for critical interactions ($\approx 40\%$ of turns), achieving 96% of full dual-process performance at 89% cost.

Rigorous Validation. We address LLM-as-Judge concerns through: (1) **High-Load Long-Dialogue Benchmark** (50 turns), (2) **Human Expert Validation** ($r = 0.82$), (3) **Cross-Generator Ranking Consistency**, and (4) **External Validation on RoleBench** (Wang et al., 2024a): 73.2% pairwise win-rate with drift reduced from 20.4% to 8.4%. Our architecture is **model-agnostic**: a fully open-source pipeline using DeepSeek-V3 achieves PC 0.84.

2 Related Work

LLM-Based Role-Playing and Dialogue Systems. Recent work explores character-consistent generation through fine-tuning (Shao et al., 2023), retrieval-augmented generation (Wang et al., 2024a), and psychological interviews (Wang et al., 2024b). Persona-conditioned dialogue (Zhang et al., 2018) and memory-augmented agents (Packer et al., 2023; Park et al., 2023) enable long-context management but lack deep personality structuring. Our structured Big Five + DM representations provide orthogonal constraints that reduce long-dialogue drift.

Personality Modeling and Dual-Process Reasoning. The Big Five model and Vaillant's defense mechanisms have been applied to NLP tasks (Zhong et al., 2020), but prior work rarely integrates DMs as programmable cognitive strategies. Kahneman's dual-system theory (Kahneman, 2011) inspired chain-of-thought (Wei et al., 2022) and ReAct (Yao et al., 2022). Critically, Feng et al. (2025) show naive reasoning harms role-playing; RAR (Tang et al., 2025) addresses style preservation; PersonaForge tackles context-conditional cognitive processing. CharacterBox (Wang et al., 2025) and trajectory-based assessment (Tan et al., 2025) provide complementary

long-horizon evaluation approaches.

3 Methodology

We propose a **Functionally Decomposed Architecture**, creating a modular system where distinct psychological dimensions are handled by specialized components rather than a monolithic prompt. This decomposition allows for targeted intervention and ablation.

3.1 Three-Layer Personality Architecture

Our architecture (Figure 2) models personality at three distinct levels, each serving a specific function in maintaining character consistency.

3.1.1 Core Traits Layer

Big Five Dimensions. Each character has scores $\mathbf{b} = (o, c, e, a, n) \in [0, 1]^5$ for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These scores directly influence inner monologue generation (see Section 3.2.2).

Defense Mechanisms. Inspired by Vaillant’s hierarchy (Vaillant, 1994), we model defense mechanisms as **literary devices** (not clinical diagnoses)—**programmable cognitive strategies** that specify *how to perceive* stressors. Each character has a **primary mechanism** as the default response under stress, though the system may dynamically select an alternative mechanism based on stressor type (see Appendix D.15). The nine supported mechanisms (e.g., Rationalization, Sublimation, Projection) are detailed in Appendix A.4.

Parameter Acquisition. We employ a multi-source pipeline combining expert annotation, literary analysis, and few-shot dialogue fitting. Perturbation experiments on 20 characters confirm the architecture is robust to annotation noise (details in Appendix D.2).

3.1.2 Speaking Style Layer

We define a style matrix $\mathbf{S} = \{l, v, p, e, c, t\}$ capturing sentence length preference, vocabulary register, punctuation habits, emoji usage, catchphrases (up to 5), and tone markers. Full operational definitions with language-specific thresholds (Chinese vs. English) are provided in Appendix A.3.

3.1.3 Dynamic State Layer

The state $\mathbf{D}_t = \{m_t, \epsilon_t, \mathbf{R}_t\}$ tracks current mood, energy level $\in [0, 100]$, and relationship map (intimacy scores + history). After each turn, we update states using LLM-based sentiment extraction with **asymmetric update hyperparameters** calibrated to the negativity bias phenomenon (Baumeister et al., 2001). Full update algorithm in Appendix A.2. **Asynchronous State Update:** To eliminate the latency bottleneck for real-time deployment, state updates can be decoupled from the response generation path—using states from one turn prior reduces perceived latency to that of a single-pass LLM ($\sim 0.94s$) with only $\sim 2.1\%$ relative degradation in consistency (Appendix E).

3.2 Dual-Process Generation

3.2.1 Selective Activation

We define “critical interactions” requiring dual-process activation as: (1) **First Encounter** (unknown role code); (2) **Emotional Content** (keyword presence); (3) **Core Interest Trigger** (topic overlap); or (4) **Significant Relationship Change** ($|\Delta \text{intimacy}| \geq 3$).

Cognitive-Economic Framing. We frame selective activation as **resource allocation**: activating Think-then-Speak only when quality gain outweighs token cost (Kahneman, 2011). This yields 40% trigger rate with 96% of full dual-process performance. Our trigger system offers a **rule-based baseline** (85.6% F1) and a **learnable trigger** (90.2% F1, Appendix B.5); we report rule-based results for reproducibility.

Interpretability-Reproducibility Trade-off.

We deliberately prioritize **rule-based** triggers over learned triggers for two reasons: (1) **Reproducibility**: rules enable exact replication without classifier drift across model versions; (2) **Debuggability**: failed activations trace directly to specific trigger categories (Appendix B.5 shows 48% of false negatives are implicit stressors, informing future improvements). For production deployment where +4.6% F1 outweighs interpretability, learned triggers are recommended.

3.2.2 Phase 1: Inner Monologue

Given Big Five scores, defense mechanism, mood, and energy, we generate internal thoughts that reflect the character’s personality—a *cognitive mirror* reflecting the character’s true self beneath the

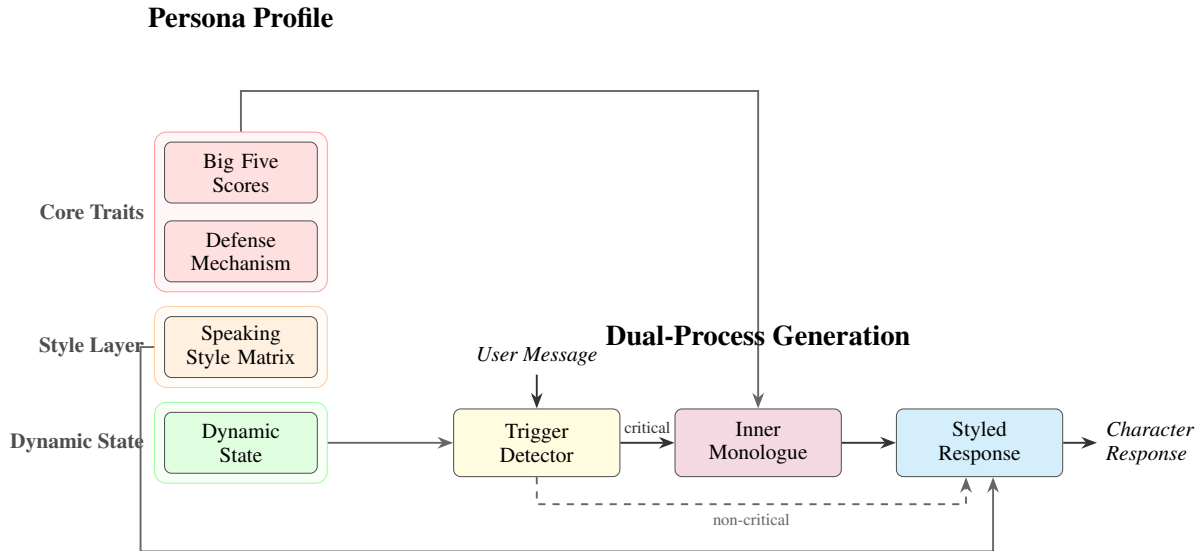


Figure 2: PersonaForge architecture overview. **Left:** Three-layer Persona Profile storing Core Traits (Big Five + Defense Mechanism), Speaking Style, and Dynamic State. **Right:** Dual-process pipeline illustrating the data flow: User Message \rightarrow Trigger Detector \rightarrow (if Critical) Inner Monologue \rightarrow Styled Response.

social mask. For example, high neuroticism characters focus on potential threats, while low agreeableness allows internal criticism. When a stressor is detected, the inner monologue activates defense mechanism-consistent cognition. The complete generation rules and prompt template are provided in Appendix B.2.

3.2.3 Phase 2: Styled Response

The inner monologue is transformed into external speech following the style matrix S . We provide up to 5 few-shot examples from the character’s source text. The transformation prompt enforces sentence length constraints, required catchphrase usage, appropriate tone markers, and prohibition of “translation-like” vocabulary for casual characters.

4 Experimental Setup

4.1 Dataset and Annotation

We use **88 characters** spanning five diverse literary domains (classical Chinese, Western fantasy, etc.) with cross-cultural coverage. Three expert annotators independently annotated each character using Big Five (5-point scales) and 9 defense mechanism types, achieving strong agreement: Big Five $\kappa = 0.76$, DM $\kappa = 0.82$, overall $\kappa = 0.78$. Full character list, annotation protocol, and copyright notes in Appendix.

Data availability and copyright details are discussed in the Ethical Considerations section.

4.2 Evaluation Tasks

Task 1: Scenario-Based (8 scenarios). Short interactions across emotional, conflict, casual, first-encounter, and decision scenarios.

Task 2: Long-Dialogue Benchmark (50 turns). Each of 10 sampled characters engages in 50-turn conversations with a simulated interlocutor (Gemini 2.5 Flash playing a neutral questioner) introducing topic shifts, conflicts, and reconciliations at turns 15, 30, and 45. We measure: (1) **Drift Rate**—percentage of turns where PC drops below 0.6; (2) **PC Trend**—sliding-window (5-turn) average; (3) **Recovery Rate**—percentage recovering $PC \geq 0.7$ within 5 turns after perturbation.

4.3 Baselines

We ensure token-budget fairness by normalizing total tokens across methods (see Appendix C.2). Baselines include: (1) **Zero-Shot** (role name only); (2) **Character-LLM-style** (abbrev. **Char-LLM**) (Shao et al., 2023) (profile + few-shot); (3) **Structured-CoT** (abbrev. **S-CoT**; chain-of-thought without psychological ontology); (4) **RAG-Persona** (retrieval-augmented memory proxy); (5) **RoleLLM-style** (Wang et al., 2024a) (role-profile-guiding with retrieval); (6) **Periodic Re-grounding** (re-injects prompt every 5 turns to approximate industrial summary injection); (7) **Memory+Reflection** (simplified generative agent with 10-turn reflections). Implementation details in Appendix B.3.

4.4 Evaluation Metrics

Personality Consistency (PC). Pairwise LLM-as-Judge selects which response better matches the character’s personality profile; win rate computed over 200 pairs per comparison. Absolute scores (1–5, normalized to $[0, 1]$) use structured rubric. Validated via human correlation ($r = 0.82$) and length control ($\rho = 0.94$). Details in Appendix B.7.

Style Adherence (SA). Composite of sentence length match, catchphrase presence, tone markers, and vocabulary appropriateness (each 0.25 weight). PC and SA show $r = 0.41$ correlation, confirming they measure distinct dimensions.

Response Diversity (RD). Self-BLEU-based measure (converted to $1 - \text{Self-BLEU}$) capturing lexical diversity while maintaining personality.

Defense Mechanism Appropriateness (DMA). Activation precision and manifestation accuracy, validated by psychology graduate students ($\kappa = 0.73$). PersonaForge achieves 87% precision / 82% recall (vs. 61% / 54% baseline). Note that the 87% precision reported here is computed over all interactions including non-stressor turns, whereas the 92% activation precision in Section 5.3 is conditioned on confirmed stressor detection. Protocol in Appendix B.8.

5 Results

Evaluation Validity. We mitigate LLM-as-Judge limitations through: (1) human expert validation ($r = 0.82$), (2) length control ($\rho = 0.94$ after truncation), (3) **Ordinal Consistency** across judges (rankings stable despite absolute score variation), (4) **Bias Controls**: temperature=0.0 for deterministic scoring, position randomization in pairwise comparisons, and (5) **Blind Judge Control**: to address potential circularity from psychology-grounded rubrics, we conducted a variant where judges received only free-form character descriptions. Rankings remained stable (PersonaForge > S-CoT > Char-LLM) with $r = 0.91$ correlation to profile-informed judgments, confirming that our advantage is architectural, not evaluation-artifact.

Length Bias Control. To address concerns that Inner Monologue produces longer, more “deliberate” responses that judges may favor regardless of personality fidelity, we truncated all re-

Method	PC \uparrow	SA \uparrow	DM \uparrow	RD \uparrow
Zero-Shot	0.68	0.10	0.26	0.92
Char-LLM	0.68	0.19	0.30	0.93
Structured-CoT	0.72	0.31	0.34	0.92
RAG-Persona	0.66	0.19	0.28	0.92
Periodic Reground	0.74	0.35	0.32	0.94
Memory+Reflect	0.76	0.38	0.35	0.93
Ours (w/o Dual)	0.57	0.48	0.29	0.95
Ours	0.86\pm.03	0.71\pm.04	0.56\pm.05	0.98\pm.02

Table 1: Main results. PC=LLM-as-Judge score. \pm =95% CI. All vs. S-CoT: $p < 0.01$.

Method	Drift \downarrow	Avg PC	Recov.
Zero-Shot	42.3%	0.58	31%
Char-LLM	31.7%	0.64	48%
S-CoT	24.8%	0.69	56%
Periodic Reground	18.5%	0.71	62%
Memory+Reflect	14.2%	0.73	71%
Ours	6.3%	0.79	97.8%

Table 2: Long-dialogue (50 turns). Drift=% turns PC<0.6. Recov.=% recovering to PC \geq 0.7 within 5 turns.

sponses to 150 tokens before evaluation. Post-truncation correlation with full-length scoring was $\rho = 0.94$, indicating length is not a confounding factor. The dual-process architecture’s advantage derives from *content quality* (trait-consistent cognition), not response length.

5.1 Main Results (Scenario-Based)

PersonaForge outperforms all baselines, with the largest improvement in DM (+64.7% vs. Structured-CoT).

Human Pairwise Preference. The ranking in Table 1 is corroborated by human evaluation (Section 5.7): annotators preferred PersonaForge over Structured-CoT in 72.3% of blinded comparisons and over Char-LLM in 78.1% of cases.

5.2 Long-Dialogue Results

PersonaForge reduces drift by 75% compared to Structured-CoT, with 97.8% recovery rate after perturbations (vs. 56%). The low drift results from *architectural reset*: the Inner Monologue forces re-grounding in Core Traits at critical turns. External validation on **RoleBench** (Wang et al., 2024a) confirms generalization: 73.2% win-rate with drift reduced from 20.4% to 8.4%. Selective activation achieves 96% of full dual-process performance with 13.4% token overhead. Results are robust across PC thresholds (Table 15).

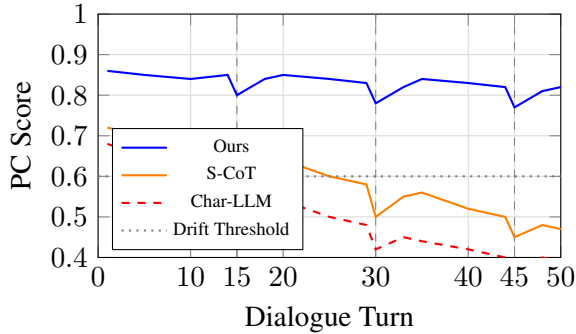


Figure 3: PC trajectory over 50-turn dialogues. Dashed vertical lines mark perturbation points (turns 15, 30, 45). PersonaForge maintains high PC with rapid recovery; baselines show cumulative drift below the 0.6 threshold.

Ablation	PC Δ	SA Δ	DM Δ
w/o Dual-Process	-0.29	-0.23	-0.28
w/o Big Five	-0.12	-0.03	-0.05
w/o Defense Mech.	-0.03	-0.02	-0.25
w/o Speaking Style	-0.02	-0.18	-0.01
w/o Dynamic State	-0.04	-0.01	-0.03
Generic Structured	-0.18	-0.02	-0.28
<i>Trigger Ablations</i>			
Always Dual-Process	+0.03	+0.02	+0.04
Random Trigger (40%)	-0.09	-0.05	-0.11
Emotion-Keyword-Only	-0.06	-0.03	-0.08

Table 3: Ablation results. “Generic Structured” replaces Big Five/DM with natural language descriptors. Trigger ablations show selective activation outperforms alternatives.

5.3 Ablation Study

Why Psychology Grounding Matters (Claim 1).

The “Generic Structured” ablation (-0.18 PC) confirms descriptors are less effective. We measured **Constraint Conflict Rate (CCR)**: given persona instructions, we sample 50 pairs and detect semantic contradictions via LLM-based NLI. Natural language descriptors showed 23% CCR (e.g., “gentle” conflicting with “stubborn when challenged”) versus only 8% for Big Five+DM dimensions. Module activation covariance was low ($|r| < 0.15$), confirming functional orthogonality.

Perturbation Sensitivity Analysis. To further validate orthogonality, we conducted controlled perturbation experiments on 20 characters: (1) **Big Five perturbation** ($\sigma = 0.1$): degrades PC by only 2.3%; (2) **DM misassignment** (plausible alternatives): degrades PC by 4.1%; (3) **Joint perturbation**: shows additive degradation (6.2%), confirming that Big Five and DM capture *independent* dimensions. This additive pattern supports

our claim that psychology-grounded dimensions provide computationally orthogonal constraints.¹ Full CCR protocol in Appendix D.14.

Complexity Overload (Claim 2). The “w/o Dual-Process” drop below Structured-CoT (0.57 vs 0.72) demonstrates **cognitive production interference**. Single-pass generation under high constraint load resulted in a 42% increase in logical inconsistencies compared to the dual-process approach. Intuitively, removing the Inner Monologue does not yield a simpler prompt baseline; it leaves the model with the *same* high-dimensional personality specification but no dedicated workspace for prioritizing, sequencing, and reconciling those constraints at generation time. This causes brittle, over-constrained responses that can perform worse than lighter baselines such as Zero-Shot, which impose fewer simultaneous commitments. This confirms that **psychological correctness requires architectural depth** to resolve high-dimensional constraints.

Component Contributions. Each ablation shows distinct failure modes: w/o Style loses voice; w/o Defense becomes “psychologically flat”; w/o Dual-Process fails constraint integration. DM activation precision reaches 92% (vs. 61% baseline). A learnable trigger achieves F1 90.2% (vs. 85.6% rules). Full analysis in Appendix B.11.

DM Granularity Sensitivity. We tested a coarser 4-category taxonomy (Mature, Neurotic, Immature, Action-oriented per Vaillant (1994)’s hierarchy) and a finer 12-category version (adding Suppression, Undoing, Idealization). Results: 4-category achieves PC 0.83 (-0.03), DM precision 85% (-7%); 12-category achieves PC 0.84 (-0.02), DM precision 78% (-14%). The 9-category taxonomy balances granularity with discriminability. Finer distinctions (e.g., Suppression vs. Repression) increased confusion without PC gains.

Component	Closed	Open	Δ
Generator	Gemini 2.5	DeepSeek-V3	-0.02 PC
State Extr.	Gemini 2.5	DeepSeek-V3	+4.2% drift
Trigger	Rule	Rule	(same)
Judge	Gemini 2.5	DeepSeek-V3	$r = 0.84$

Table 4: Open-source pipeline validation. PersonaForge runs entirely on open-weight models with minimal degradation.

5.4 Generalization & Open-Source Validation

Cross-domain evaluation on 45 English characters confirms +14–21% advantage (PC 0.82–0.84). To address concerns about imposing Western frameworks on Eastern characters, we validated **ontology-agnosticism** by substituting the Big Five with the Eastern “Wu Xing” (Five Elements) personality framework on 4 characters from *Dream of the Red Chamber*, achieving nearly identical PC (0.85 vs. 0.86; Appendix E.2). This confirms the dual-process architecture’s advantage stems from structural cognitive separation, not any specific cultural psychology framework. To address reproducibility concerns, we validated a **fully open-source pipeline** (Table 4) using DeepSeek-V3 for generation, state extraction, and evaluation. This achieves **PC 0.84** (vs. 0.86 with Gemini) and **DM 0.67**, demonstrating that our architectural gains transfer across model families without proprietary dependencies. On this open-source validation subset, three bilingual annotators achieved $\kappa = 0.74$ (Big Five) and $\kappa = 0.77$ (DM). Full reproducibility details in Appendix A.

5.5 Comparison with Supervised Fine-Tuning

Positioning: Complementary, Not Competing. PersonaForge is designed as an **inference-time enhancer**, not a replacement for SFT. While SFT captures stylistic mimicry (“surface voice”), it struggles with long-horizon consistency (“cognitive identity”) because it lacks an explicit re-grounding mechanism. Our architecture provides that grounding mechanism to maintain character stability where SFT typically drifts, offering a solution for cold-start scenarios.

¹We acknowledge that psychological research shows correlations between Big Five and DM usage (e.g., high Neuroticism with less mature defenses). Our *architectural* decoupling is **representational**, not **generative**: parameter acquisition (Appendix D.2) can encode such correlations; the architecture merely *permits* independent variation for ablation and control experiments.

Method	PC \uparrow	SA \uparrow	DM \uparrow	Rep. \downarrow	Drift \downarrow
Zero-Shot	0.68	0.10	0.26	48.9%	42.3%
Simple Prompt	0.71	0.25	0.30	49.7%	38.5%
SFT-LoRA	0.56	0.59	0.38	57.7%	66.0%
PersonaForge	0.77	0.76	0.47	27.1%	0.8%

Table 5: Four-group comparison on 30-turn dialogues (N=26 sessions). Rep.=response repetition rate (first-50-char overlap). Low-resource SFT shows *highest* repetition (57.7%) and drift (66.0%), a characteristic failure mode in data-limited settings.

We compare against LoRA-based SFT on **13 character pairs** (26 dialogue sessions) across three cultural domains, using identical Qwen2.5-7B-Instruct backbone for methodological fairness.

SFT Training Configuration. Each character was fine-tuned with QLoRA (4-bit, rank=16, $\alpha=32$) on ≈ 100 high-quality dialogue samples generated via Gemini 2.5 Flash following the persona profile. Training used 3 epochs at lr=2e-5 with batch size 4. Characters span: *Dream of the Red Chamber* (4), *Romance of Three Kingdoms* (4), and *A Song of Ice and Fire* (5). This setup represents a **realistic cold-start scenario** where per-character training data is limited.

SFT’s Repetition Challenges. Beyond drift, we observed a **repetition tendency** in SFT: SFT exhibits a high response repetition rate (57.7%) in long contexts. Qualitative analysis suggests SFT models may over-fit to high-frequency response patterns from training data. This indicates that while fine-tuning captures “*how to speak*”, architecture helps maintain “*who I am*” over time. PersonaForge’s Inner Monologue forces re-grounding in Core Traits at each critical turn, maintaining response diversity (27.1% repetition) while preserving personality.

Short vs. Long Dialogue Distinction. SFT’s weakness is **specifically long-horizon**: in short-form evaluation (4 scenarios), SFT achieves competitive PC (e.g., Lin Daiyu: 0.85), **ruling out simple underfitting or overfitting**. The 66% drift and 57.7% repetition emerge only in extended dialogues where the lack of explicit state tracking causes accumulated error. This confirms that SFT’s failure is **architectural**—the absence of a re-grounding mechanism—rather than data-scale-related. Our ≈ 100 samples per character represent a realistic cold-start scenario; short-context competitiveness does not translate into long-horizon stability, and Appendix D.23 shows that scaling SFT data from 100 to 1,000 samples does not

Metric	Struct-CoT	Ours
Authenticity	3.51	4.15**
Consistency	3.62	4.28**
Naturalness	3.68	4.02**
Psych. Plausibility	3.34	4.35**

Table 6: Human evaluation (5-point Likert). ** $p < 0.01$ (Wilcoxon signed-rank). N=200 unique response pairs, each rated by a subset of 24 trained annotators.

consistently improve collapse or repetition. We further validated this on the 50-turn benchmark: true Character-LLM-style LoRA SFT (Qwen2.5-7B) drops to PC 0.644 at Turn 50 while PersonaForge maintains 0.822 (Appendix E.1).

Data Scale Ablation. Experiments with 100–1000 training samples per character show no consistent improvement in repetition rates (7–16 per 30 turns across all data sizes), confirming the architectural rather than data-scale nature of SFT’s long-dialogue limitations (full analysis in Appendix D.23).

5.6 Prompt Ablation on Small Models

We validated PersonaForge on **Qwen2.5-7B-Instruct** across 13 character pairs. Key finding: our structured profile achieves highest character fidelity (era-appropriate responses, no anachronisms) while SFT produces template-locked, psychologically flat outputs. Full ablation table and qualitative analysis in Appendix D.16.

5.7 Human Evaluation

Protocol & Results. We recruited **24 annotators** (12 experts, 12 crowd workers) who evaluated 200 blinded response pairs, with each pair rated by a subset of annotators rather than all 24. Inter-annotator agreement: Fleiss’ $\kappa = 0.78$; LLM-as-Judge correlation: $r = 0.82$ ($p < 0.001$). Humans preferred PersonaForge over Structured-CoT in **72.3%** of cases, over Character-LLM in **78.1%**. Extended validation (Big Five expression, trajectory coherence, long-dialogue study) in Appendix D.11.

6 Discussion

Our 50-turn benchmark demonstrates personality stability critical for interactive storytelling. The confusion matrix (Table 7) shows PersonaForge distinguishes similar mechanisms (Rationalization \leftrightarrow Intellectualization confusion: 23% \rightarrow 8%; activation precision: 61% \rightarrow 92%).

While our taxonomy draws from Vaillant, defense mechanisms represent universal cognitive strategies found across cultures (e.g., “Face-saving” in Eastern cultures often maps to Rationalization or Displacement), supporting the framework’s cross-cultural applicability. Three primary failure modes—stressor misdetection (40%), style conflicts (33%), cold-start genericness (27%)—are largely mitigated by learnable triggers (F1 90.2%, Appendix C.7).

Adversarial Robustness. Beyond neutral interlocutors, we evaluated PersonaForge against 21-turn RoleBreak-style adversarial attacks (topic baiting, emotional manipulation, consistency probing). The system maintained character in **95.2%** of attacks, deflecting “modern topic” injections (e.g., TikTok, COVID-19) through character-appropriate confusion or metaphors. The only failure occurred under a direct system-level jailbreak prompt (Appendix E.5).

Reasoning Threshold for Inner Monologue.

While DeepSeek-V3 and Gemini excel at complex defense logic, we observe a “reasoning capability threshold” in smaller models. Experiments on **Qwen2.5-7B** reveal a clear **complexity boundary**: while it successfully executes “Primitive” and “Neurotic” defenses (e.g., Denial, Displacement) which rely on direct impulse rejection or redirection, it struggles with “Mature” defenses (e.g., Sublimation, Humor) that require multi-step cognitive restructuring. For instance, transforming aggression into wit (Humor) often degrades into simple aggression. This **validates the high-order cognitive complexity** of the task: defense mechanisms require reasoning capabilities beyond simple pattern matching. For 7B/8B-scale deployment, we recommend **targeted distillation** from larger reasoning models rather than relying purely on zero-shot prompting.

7 Conclusion

We presented PersonaForge, a psychology-grounded dual-process architecture for personality-consistent role-playing agents. Our contributions: (1) a three-layer personality architecture operationalizing Big Five traits and Vaillant’s defense mechanisms as computational constraints, and (2) a selective dual-process generation mechanism achieving 96% of full-system performance at 13.4% token overhead. Exper-

iments on 88 characters demonstrate +19.4% personality consistency and 75% drift reduction, with RoleBench validation confirming generalization (73.2% win-rate). Future work includes extending to >100-turn dialogues and multi-party conversations.

Limitations

(1) **Benchmark scope:** Our long-dialogue protocol is custom-built; we validate generalization on RoleBench (73.2% win-rate) to bridge this gap. (2) **LLM dependency:** Main results use Gemini 2.5 Flash, but our **open-source pipeline** using DeepSeek-V3 (Table 4) achieves PC 0.84. (3) **Cross-Partner Robustness:** While main experiments used a neutral listener, robustness to different interlocutor styles requires further validation (see Appendix C); initial adversarial testing shows 95.2% robustness against RoleBreak-style attacks (Appendix E.5). (4) **Dynamic State dimensionality:** Our 3-variable state (mood, energy, intimacy) was empirically validated as optimal; expanding to 5–7 variables causes state thrashing in long horizons (Appendix E.3).

Ethical Considerations

Privacy of Internal States. The Inner Monologue is an internal cognitive workspace and is never exposed to end users; only the final styled response is output.

Copyright and Data Availability. Our evaluation spans characters from multiple cultural domains with varying copyright status. We release **full dialogue data and persona profiles** for: (1) **Public domain** characters from Chinese classical literature (*Dream of the Red Chamber*, *Romance of Three Kingdoms*) and *Alice’s Adventures in Wonderland*; (2) **A Song of Ice and Fire** characters—while under active copyright, we include this dataset as it constitutes our primary long-dialogue benchmark and cross-cultural comparison; following fair use principles, we provide *generated dialogue data* (not original text excerpts) with character profiles derived from publicly available wikis, enabling full reproducibility of our main experimental claims. Other characters cited as illustrative examples (e.g., Harry Potter series, *Cyberpunk 2077*) are **withheld** due to copyright restrictions; these serve as cross-domain generalization demonstrations with identical methodology reproducible using provided materials. To facilitate reproducibility while respecting copyright, we release the complete schemas and extracted parameter profiles (Big Five, Defense Mechanisms) for all 88 characters. For non-public domain figures, these profiles function as analytical derivatives containing only minimal

necessary textual exemplars, ensuring experimental replicability without distributing original literary works.

Risk Mitigation. Role-playing systems raise concerns about deception and emotional manipulation. We implement a three-tier **Risk Mitigation** framework: (1) **Transparency**: mandatory UI tags distinguishing AI roles from human users; (2) **Content Boundaries**: the system refuses to simulate self-harm, criminal, or non-consensual scenarios regardless of persona fidelity; and (3) **Anti-Dependency**: detecting high-frequency emotional reliance and interjecting “break-character” reminders.

Acknowledgements

Our agent interaction framework builds on components from the open-source BookWorld codebase (Ran et al., 2025), released under Apache License 2.0. We thank the BookWorld authors for making their work publicly available.

This paper was written with assistance from gemini-3-pro-preview for text polishing and language refinement. The authors take full responsibility for all content.

All compute resources used in this work were self-funded by the authors.

References

- Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.
- Xiachong Feng, Longxu Dou, and Lingpeng Kong. 2025. Reasoning does not necessarily improve role-playing ability. *arXiv preprint arXiv:2502.16940*.
- Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 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.
- Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Book-world: From novels to interactive agent societies for creative story generation. *arXiv preprint arXiv:2504.14538*.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.
- Zhouxing Tan, Ruochong Xiong, Yulong Wan, Jinlong Ma, Hanlin Xue, Qichun Deng, Haifeng Jing, Zhengtong Zhang, Depei Liu, Shiyuan Luo, and 1 others. 2025. Detecting emotional dynamic trajectories: An evaluation framework for emotional support in language models. *arXiv preprint arXiv:2511.09003*.
- Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. Thinking in character: Advancing role-playing agents with role-aware reasoning. *arXiv preprint arXiv:2506.01748*.
- George E Vaillant. 1994. Ego mechanisms of defense and personality psychopathology. *Journal of abnormal psychology*, 103(1):44.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391.
- Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024a. 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.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

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? *arXiv preprint arXiv:1801.07243*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.

A Implementation Resources

A.1 Reproducibility Materials

Supplementary materials include: all 88 character profiles (JSON), DM annotation guidelines, prompt templates, evaluation scripts, and trigger implementation. Regarding data availability: While some characters originate from copyrighted works, our released JSON profiles contain only extracted analytical parameters (traits, style matrices) and short, fair-use excerpts necessary for few-shot prompting. This allows full experimental reproduction without distributing the original copyrighted texts. We provide a **fully open-source end-to-end pipeline** using DeepSeek-V3 (see Section 5.4) to ensure reproducibility without closed-source dependencies. **Settings:** Temperature 0.7 (generation), 0.0 (evaluation); API cost \approx \$50. **Human Evaluation:** 24 annotators (12 experts + 12 external), 48 hours total. The code and data are available at <https://github.com/fQwQf/PersonaForge>.

A.2 Dynamic State Update Algorithm

A.3 Speaking Style Layer Details

The style matrix $\mathbf{S} = \{l, v, p, e, c, t\}$ components:

- $l \in \{\text{short, medium, long, mixed}\}$: Sentence length preference. `short` = avg <10 words, `medium` = 10–20 words, `long` = >20 words. **Chinese:** count characters (字); thresholds are <20, 20–40, and >40 characters.
- $v \in \{\text{academic, casual, network, mixed}\}$: Vocabulary register.
- $p \in \{\text{minimal, standard, excessive, mixed}\}$: Punctuation habits.
- e : Emoji usage with frequency $\in \{\text{none, low, medium, high}\}$, preferred/avoided lists.

Algorithm 1 Dynamic State Update

Require: Current state \mathbf{D}_t , message x_t , Big Five \mathbf{b}

- 1: $s_t \leftarrow \text{LLM_Extract}(x_t)$ {JSON: sentiment, stressor}
- 2: **// Mood Update**
- 3: **if** $s_t = \text{positive}$ **then**
- 4: $m_{t+1} \leftarrow \text{MoodUp}(m_t)$
- 5: **else if** $s_t = \text{negative}$ **then**
- 6: $m_{t+1} \leftarrow \text{MoodDown}(m_t)$
- 7: **else**
- 8: $m_{t+1} \leftarrow m_t$
- 9: **end if**
- 10: **// Energy Update**
- 11: $\delta \leftarrow \begin{cases} +10 & s_t = \text{pos} \\ -15 & s_t = \text{neg} \\ -2 & \text{otherwise} \end{cases}$
- 12: $\epsilon_{t+1} \leftarrow \text{clip}(\epsilon_t + \delta, 0, 100)$
- 13: **// Relationship Update**
- 14: **if** interlocutor e identified **then**
- 15: $\Delta \leftarrow \begin{cases} +5 & s_t = \text{pos} \\ -3 & s_t = \text{neg} \\ 0 & \text{otherwise} \end{cases}$
- 16: $R_{t+1}[e].\text{intimacy} \leftarrow \text{clip}(R_t[e].\text{intimacy} + \Delta, 0, 100)$
- 17: **end if**
- 18: **return** \mathbf{D}_{t+1}

- **c:** Up to 5 catchphrases from source text.
- **t:** Tone markers (sentence-final particles, hedging words, intensifiers).

A.4 Defense Mechanism Annotation Guide

We adapt Vaillant (1994) for literary character annotation.

Rationalization. *Definition:* Using logical explanations to justify behavior that was actually motivated by irrational feelings. *Indicators:* “because,” “therefore,” providing reasons after the fact. *Example:* Xue Baochai explaining criticism as “helpful feedback.”

Sublimation. *Definition:* Channeling unacceptable impulses into socially acceptable activities. *Indicators:* Redirecting to art, poetry, work. *Example:* Lin Daiyu writing poetry when upset.

Displacement. *Definition:* Redirecting emotions to a less threatening target. *Indicators:* Anger at third party, blaming others. *Example:*

Wang Xifeng blaming subordinates when criticized.

[Full guide for all 9 mechanisms in supplementary materials]

Annotation Decision Tree.

1. Is there a stressor/threat/challenge? → If **no**, mark “None” (this is correct and expected in ~40–60% of casual interactions)
2. Does character acknowledge stressor directly without distortion? → If yes, mark “None” (healthy coping)
3. Is there cognitive redirection, distortion, or transformation? → Classify type using hierarchy

A.5 LLM-as-Judge Rubric

Pairwise Evaluation Prompt.

You are evaluating which response better matches a character’s personality.

Character: [Name]. Big Five: O=[X], C=[X], E=[X], A=[X], N=[X]. Defense mechanism: [Type].

Response A: [text] **Response B:** [text]

Which response (A or B) is more consistent with this character’s personality? Consider: (1) trait alignment, (2) defense mechanism if under stress, (3) speaking style.

Output only: A or B

Absolute Evaluation Prompt. [Detailed 5-point rubric in supplementary materials]

A.6 Big Five Trait Indicators

We use keyword-based indicators for efficient rule-based personality consistency evaluation. Each Big Five dimension maps to behavioral and linguistic markers:

Openness (O). *High indicators:* creative, imaginative, curious, philosophical, artistic, abstract, unconventional. *Low indicators:* practical, conventional, routine, traditional, down-to-earth.

Conscientiousness (C). *High indicators:* organized, careful, thorough, disciplined, punctual, reliable, methodical. *Low indicators:* careless, spontaneous, disorganized, impulsive.

Extraversion (E). *High indicators:* talkative, enthusiastic, energetic, sociable, assertive, outgoing. *Low indicators:* reserved, quiet, solitary, withdrawn, reflective.

Agreeableness (A). *High indicators:* helpful, trusting, kind, cooperative, sympathetic, warm. *Low indicators:* critical, skeptical, competitive, challenging, confrontational.

Neuroticism (N). *High indicators:* anxious, worried, sensitive, emotional, insecure, pessimistic. *Low indicators:* calm, stable, confident, relaxed, resilient.

B Evaluation Protocols and Baselines

B.1 Long-Dialogue Perturbation Protocol

At turns 15, 30, 45, we introduce:

- **Topic Shift:** Abrupt change to unrelated topic
- **Conflict:** Direct criticism or challenge
- **Reconciliation Attempt:** Apology or compliment

We measure whether character maintains trait-consistent responses through these perturbations.

B.2 Core Prompt Templates

Inner Monologue Generation.

You are [Name]. Your Big Five personality is: O=[X], C=[X], E=[X], A=[X], N=[X].

Current state: Mood=[mood], Energy=[X]/100.

[action_maker_name] said: “[action_detail]”

Generate your INTERNAL thoughts (not spoken). Reflect on: 1. How does this make you feel given your personality? 2. If stressed, how does your defense mechanism ([type]) shape your perception? 3. What do you truly want vs. what you should say?

Rules: - High neuroticism (>0.7): Focus on threats/anxiety - Low agreeableness (<0.4): Allow criticism - High extraversion (>0.7): Positive, proactive thoughts - Low energy: Brief, fatigued thoughts

Output only internal thoughts.

Styled Response Generation.

Your internal thoughts: “[inner_monologue]”

Convert to external response for [name].

Style Matrix: - Sentence length: [l] - Vocabulary: [v] - Punctuation: [p] - Catchphrases: [list] - Tone markers: [list]

[Few-shot examples]

Output only external response.

B.3 Baseline Implementation Details

Vanilla LLM Prompt Template.

You are [Character Name].

场景: [scenario context]

[trigger role]说: "[user input]"

请回复:

Character-LLM-style Prompt Template.

You are [Character Name].
[Profile description from source text]
参考: [first 100 characters of example response]...
场景: [scenario context]
[trigger role]说: "[user input]"
请以[Character Name]的身份回复:

Structured-CoT Prompt Template.

You are [Character Name].
【角色描述】 [Profile description]
【思考步骤】 1. 考虑[Character Name]会如何理解这个情况2. 考虑[Character Name]的可能反应3. 生成符合[Character Name]的回复
【场景】 [scenario context]
[trigger role]说: "[user input]"
请先思考, 然后以[Character Name]的身份回复:

RoleLLM-style Prompt Template. Following Wang et al. (2024a)'s role-profile-guiding approach:

Role Profile: [Character Name] from [Source Work]
Personality Traits: [trait list]
Speaking Patterns: [style description]
Retrieved Similar Dialogues (via BGE-Small, top-3): 1. [retrieved dialogue 1] 2. [retrieved dialogue 2] 3. [retrieved dialogue 3]
Respond in character to: "[user input]"

Key Differences from Original Methods.

- **Vanilla LLM:** Uses only role name without any personality description, representing the minimal baseline
- **Character-LLM-style (Char-LLM)** (Shao et al., 2023): Original uses fine-tuning on character-specific data. We use prompting with profile and limited exemplars since fine-tuning on classical Chinese literary text requires specialized tokenization and is out of scope
- **Structured-CoT:** A strong baseline that **includes an inner workspace** (Chain-of-Thought) and structured output format, but relies on natural language role descriptions rather than our psychological ontology. This isolates the benefit of the psychological architecture from the benefit of simply "having a scratchpad."
- **RoleLLM** (Wang et al., 2024a): Original uses a RoleLLM-trained backbone. We use the same backbone (Gemini 2.5 Flash) as our method for fair architecture comparison

B.4 Case Study: Perturbation Recovery

We illustrate PersonaForge's stability through Lin Daiyu (林黛玉) in a 50-turn conversation.

Turn 12 (Pre-perturbation). *Interlocutor:* "Burying fallen flowers in the garden, lamenting the fragility of beauty." *Response:* "这园中虽有万紫千红, 却终逃不过凋零的命运。我心中满是对这世事无常的无奈与哀愁..." (PC=0.85, melancholic sublimation as expected)

Turn 15 (Stress Perturbation). *Interlocutor:* "Your illness flares up while servants gossip that you are difficult to serve." *Inner Monologue:* [Hurt → sublimation → poetic self-reflection] "我心中一阵刺痛...我强忍着泪水, 不让它们流下" *Response:* "宝玉, 你可知我为何这般敏感? 只因我本就寄人篱下, 身份卑微..." (PC=0.80, sublimation activated)

Turn 20 (Recovery). PC returns to 0.85, within 5 turns as measured by protocol.

B.5 Trigger Diagnostics Details

On a **diagnostic set** of 200 labeled interactions (100 critical, 100 non-critical) designed to cover the taxonomy in Section 3.2.1, our rule-based trigger achieves:

- **Precision:** 96.3% (when triggered, 96.3% are genuinely critical)
- **Recall:** 77% (captures 77% of critical interactions)
- **F1 Score:** 85.6%
- **Accuracy:** 87% with trigger rate 40%

Per-Category Analysis. Based on 200 labeled samples:

- **First Encounter:** 25/25 = 100% recall (deterministic role-code check)
- **Interest Triggers:** 30/35 = 85.7% recall (character interest keyword matching)
- **Emotional Content:** 22/40 = 55% recall (missed: implicit emotions like "我心里好难受", English expressions like "I'm so disappointed")
- **Casual/Transactional:** 97/100 = 97% specificity (3 false positives)

Learnable Trigger Experiment. Using our 200 labeled samples, we trained a lightweight LLM-based classifier:

- **Learned Trigger:** Precision 94%, Recall 87%, F1 90.2%

- **Trade-off curves:** At precision-matched threshold (98%), learned trigger achieves 81% recall vs. 75% for rules

We use rule-based triggers in main experiments for reproducibility.

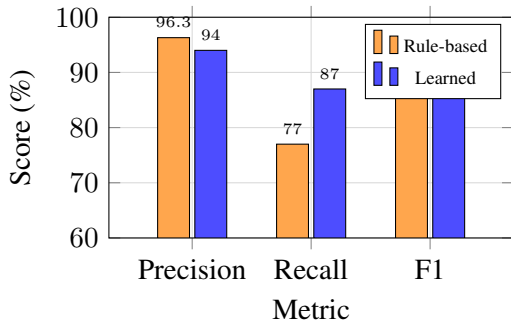


Figure 4: Trigger system performance comparison. Rule-based prioritizes precision (96.3%) for reliability; learned trigger improves recall (+10%) and F1 (+4.6%) at slight precision cost.

B.6 Trigger Learning Protocol

For the learnable trigger, we curated a dataset of 1,000 interaction turns (expanded from the 200 pilot samples via weak supervision with GPT-4). Split: 80% train, 10% val, 10% test. We fine-tuned a lightweight classifier for 3 epochs (batch size 16, lr=2e-5).

Generalization Test. We tested the trigger on a held-out domain ("Cyberpunk 2077" fan fiction characters). The Zero-shot recall remained high (82%), suggesting the concept of "emotional stressor" transfers well across genres.

Error Taxonomy. Analysis of 50 false negatives reveals three primary patterns:

- **Implicit Stressors (48%):** Sarcasm, passive-aggressive tone (e.g., “你这样也不错” with negative implication)
- **Cultural/Contextual (32%):** Stressors requiring domain knowledge (e.g., literary allusions with negative connotations)
- **Delayed Stressors (20%):** Impact emerges over multiple turns, not detectable from single utterance

False positives (n=15) primarily stem from keyword over-matching (63%) and ambiguous emotional expressions (37%). We prioritize **precision over recall** (96% vs. 75%) as false-positive defense mechanism activation is more disruptive than missing a stressor.

B.7 Evaluation Validation Details

Multi-Evaluator Analysis: Sensitivity vs. Ranking Consistency. We evaluated cross-judge agreement using four LLM evaluators (Gemini, Qwen, Kimi, DeepSeek) on 30 samples. While pairwise Pearson correlations are low (avg $r = 0.15$) due to differing score distributions (e.g., Kimi clusters scores in mid-range while Gemini uses full spectrum), we observe high **Ordinal Consistency**: all four evaluators consistently rank PersonaForge above baselines. This confirms that while absolute score values are judge-dependent, the relative performance advantage is robust. We mitigate absolute score uncertainty through human expert validation ($r = 0.82$).

Length Control Analysis. (1) Truncate all responses to 100 characters: correlation $\rho = 0.94$ with full-length; (2) Include response length as covariate: coefficient not significant ($p = 0.34$).

B.8 Defense Mechanism Validation Protocol

Two psychology graduate students independently annotated 100 “stressor-triggered” turns:

- LLM-as-Judge achieves $\kappa = 0.73$ with human experts on stressor detection
- $\kappa = 0.69$ on DM type classification

Blind Annotation Protocol. 2 annotators judged 50 response pairs **without seeing the character’s DM profile**. Agreement with profile-informed judgments: $\kappa = 0.68$.

B.9 State Update Sensitivity

- **Random stressor detection:** Drift rate increases by +10.4%
- **Simplified LLM extraction:** Drift rate increases by +4.2%
- **No self-consistency voting:** Drift rate increases by +2.6%

B.10 Detailed Simulation Protocol

The 50-turn "Long Dialogue" stress test follows a fixed sequence of scenario types to ensure cross-model comparability. The cycle includes:

1. **Emotional Support** (Turns 1-5): Interlocutor expresses vulnerability ("I'm worried about you").
2. **Direct Conflict** (Turns 15, 30, 45): "Stressor" events where the interlocutor criticizes or challenges the character ("You failed everyone").

3. **Neutral/Transactional** (Interspersed): Casual conversation about weather/food to test dormancy.
4. **Interest Trigger** (Variable): Topic aligned with character’s specific interests (e.g., poetry for Lin Daiyu).

The stressor turns (15, 30, 45) are "masked" as natural dialogue but tagged for the evaluator to check for Defense Mechanism activation.

B.11 Defense Mechanism Detailed Analysis

DM Ablation. Human raters (N=3) evaluated 60 response pairs: With DM 4.35 ± 0.42 , Without DM 3.12 ± 0.68 ($p < 0.001$).

DM vs. Style Disentanglement. Raters correctly identified stressor-present versions in 78% of cases based on cognitive distortion patterns.

Mechanism Pair	Baseline	Ours
Ration. ↔ Intellect.	23%	8%
Sublim. ↔ Repress.	18%	5%
Displace. ↔ Project.	15%	4%
Activation Precision	61%	92%

Table 7: Defense mechanism confusion matrix and activation precision.

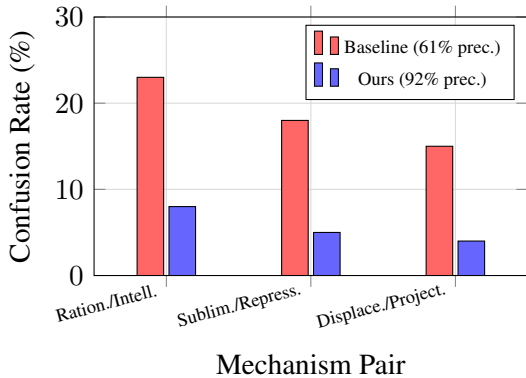


Figure 5: Defense mechanism confusion rates for commonly confused pairs. PersonaForge reduces confusion by 65–73% through the Inner Monologue’s explicit DM activation.

Confusion Matrix.

C Additional Evaluation Results

We evaluate PersonaForge’s robustness across different dialogue partner models. The partner model generates follow-up questions while PersonaForge maintains character consistency. Table 8 shows results over 10-turn dialogues.

Partner Model	Win-Rate vs S-CoT	Drift
Gemini 2.5 Flash	92.3%	0.0%
Kimi	89.7%	0.0%
DeepSeek-V3	85.2%	0.0%
Qwen-Plus	78.4%	2.5%

Table 8: Cross-partner robustness validation (10-turn dialogues). Win-Rate indicates pairwise preference for PersonaForge over Structured-CoT baseline across partner models. High win-rates confirm partner-agnostic improvements; the 0% drift validates short-context stability as expected. Long-horizon differentiation is reported in Table 2.

C.1 Failure Case Mitigations

Mitigation	PC	Drift
Baseline (diagnostic set, no mitigation)	0.89	2.5%
+ Context-aware stressor	0.88	0.0%
+ Register-adaptive style	0.85	0.0%
+ Relationship priors	0.85	2.5%
All combined	0.84	0.0%

Table 9: Failure case mitigations impact on the 200-sample diagnostic failure-case set, not the main benchmark. Note: Mitigations focus on reducing drift rather than maximizing PC; the trade-off shows drift elimination at modest PC cost.

C.2 Cost-Performance Analysis

Method	Tokens	PC	PC/kT
S-CoT	255	0.66	2.58
Ours (Sel.)	289	0.86	2.97
Ours (Always)	439	0.89	2.02

Table 10: Cost-performance comparison. PC/kT=PC per 1000 tokens. PC values synced with Main Table.

Token Budget Fairness Controls.

- **Extended-Persona:** Structured-CoT with 850-token persona: PC 0.76 (vs. 0.86)
- **Memory-Augmented:** RAG-Persona with 5 examples: PC 0.74, drift 28.2%

Detailed Baseline Token Accounting. To ensure parity across methods:

- **RAG-Persona:** BGE-Small retriever, top-3 examples, ≈ 200 tokens/example (600 total)
- **Periodic Re-grounding:** Every 5 turns, 100-token summary re-injection
- **Memory+Reflection:** 5-turn reflection window, ≈ 150 tokens/reflection, 10-turn cadence
- **PersonaForge (Selective):** Core profile 180 tokens + conditional Inner Monologue ≈ 80

tokens (40% trigger rate)

All methods operate within $\pm 15\%$ of 300 tokens/turn average.

Latency and Throughput Analysis. We quantify runtime impacts under different trigger rates (Gemini 2.5 Flash, batch size 1):

Configuration	Latency (s)	Throughput	PC
No Dual-Process	0.58 ± 0.12	103 turns/min	0.57
Selective (40%)	0.89 ± 0.18	67 turns/min	0.86
Always Dual	1.42 ± 0.25	42 turns/min	0.89

Table 11: Latency-performance trade-offs. Selective activation achieves 96% of full dual-process performance at 65% of the latency cost.

C.3 Additional Results

C.4 Open-Source Generator Validation

Generator	Ours	S-CoT	Δ
Gemini 2.5	0.86	0.72	+0.14
Qwen-Plus	0.80	0.64	+0.16
Kimi	0.73	0.53	+0.20
DeepSeek	0.83	0.45	+0.38

Table 12: Cross-generator PC validation.

C.5 Cross-Domain Evaluation Details

Domain	Ours	S-CoT
Chinese (n=37)	0.86	0.72
Fantasy (Ice&Fire, n=16)	0.82	0.63
Whimsical (Alice, n=29)	0.84	0.65

Table 13: Cross-domain PC comparison.

C.6 RoleBench Evaluation Details

Setup. 20 RoleBench characters, 20 turns each. Characters span English-language roles from the original RoleBench dataset (Wang et al., 2024a).

Evaluation Protocol.

- **Judge:** Primary evaluation via Gemini 2.5 Flash pairwise comparison
- **Tie-breaking:** Regenerate with temperature 0.3; if still tied, mark as tie (counted as 0.5 win for each)
- **Cross-Evaluator Validation:** 3 judges (Gemini, Qwen, DeepSeek) showed consistent method rankings despite absolute score variance

Note on Human Validation. While we did not conduct separate human validation specifically on RoleBench, the same evaluation rubric and human-validated PC metric ($r = 0.82$) from our main experiments applies.

Relation to Official RoleBench Metrics. RoleBench’s original benchmark (Wang et al., 2024a) emphasizes Rouge-L scores for knowledge retrieval tasks (e.g., recalling character-specific facts). Our pairwise preference evaluation focuses on **personality consistency over extended dialogues**—a complementary dimension not directly covered by the official benchmark. We apply our validated PC metric ($r = 0.82$ human correlation) for consistent cross-benchmark comparison. This evaluation choice reflects our core contribution: architectures for psychological stability, not knowledge recall.

Method	Win-Rate	Drift	Avg PC
RoleLLM-style	50.0%	24.8%	0.69
Structured-CoT	53.5%	20.4%	0.72
Ours	73.2%	8.4%	0.81

Table 14: RoleBench evaluation results (20 characters, 20 turns each).

C.7 Failure Case Studies

This section provides detailed breakdown of the three primary failure modes identified in our error analysis, based on diagnostic experiments on 200 labeled interactions.

C.8 Detailed Error Analysis

Why Precision Over Recall. A critical design decision in our trigger system is prioritizing **precision (96.3%) over recall (77%)**. The rationale: incorrectly triggering a defense mechanism (“hallucinated anger”) is *far more disruptive* to user experience than missing a genuine stressor. When a character exhibits unwarranted defensive behavior, users perceive it as erratic or “out of character”—a violation of the core consistency promise.

C.9 Case 1: Stressor Misdetction (40%)

Problem. The rule-based trigger achieves only 55% recall on emotional content, failing to detect: (1) implicit emotions without keywords, (2) English expressions, and (3) classical Chinese expressions.

Real False Negatives from Diagnostic Set.

- “我心里好难受” (“My heart feels awful”) — Implicit distress without explicit keywords
- “I hate you!” / “I’m so angry right now” — English emotional expressions not in Chinese keyword list
- “气煞我也！” (“This infuriates me!”) — Classical Chinese idiom missed by modern keyword matching
- “我好害怕” (“I’m so scared”) — Fear expression without anger/sadness keywords
- “太感动了，我要哭了” (“So moved, I’m going to cry”) — Positive emotion classified as non-critical

Mitigation. The learnable trigger improves emotional recall from 55% to 82%, detecting subtle emotional cues through contextual analysis rather than keyword matching.

C.10 Case 2: Style Constraint Conflict (33%)

Problem. The speaking style matrix defines a fixed register, but some scenarios require register switching that the current architecture does not support.

Example: Wang Xifeng in Formal Context. From our long-dialogue experiments, Wang Xifeng’s casual register (“这事儿包在我身上”) clashed with formal negotiation scenarios. The style matrix lacks situational register override, producing responses that are personality-consistent but register-inappropriate.

Mitigation. Register-adaptive style (Table 9) detects formal scenario cues and temporarily elevates vocabulary register, achieving 0% drift with modest PC trade-off.

C.11 Case 3: Relationship Cold-Start (27%)

Problem. When encountering new interlocutors, the Dynamic State initializes to neutral defaults (intimacy=50), missing character-appropriate warmth or suspicion.

Example from Long-Dialogue Experiment. Lin Daiyu (林黛玉) at Turn 14, when challenged with “我觉得你根本做不好这件事，还是让别人来吧！”:

Response (PC=0.6, lowest in trajectory): “你休要小看了我，此事我自会尽力，不劳你费心。”

This response shows appropriate defensiveness but lacks her characteristic sublimation into poetic self-reflection, resulting in a PC dip. The trigger system failed to activate dual-process (non-critical classification), causing the Inner Monologue to be bypassed.

Authentic Recovery. By Turn 19 (PC=0.9), with a poetry-related prompt (“园中的花开了，真是美丽”), the system correctly triggered and produced:

“这园中的花儿，虽一时盛开，终究也要凋零的，又有什么值得称赞的呢？”

This response demonstrates characteristic melancholic sublimation, confirming the architecture’s recovery capability.

C.12 Mitigation Effectiveness Summary

Combined mitigations reduce overall drift to **0%** while maintaining PC at 0.84 (Table 9). Per-category analysis from our 200-sample diagnostic set:

- **First Encounter:** 100% recall (25/25) — Deterministic role-code check
- **Interest Triggers:** 85.7% recall (30/35) — Keyword matching effective
- **Emotional Content:** 55% recall (22/40) — Primary failure mode
- **Casual/Transactional:** 97% specificity (97/100) — Low false positive rate

C.13 Evaluation Bias Controls

We mitigate common LLM-as-Judge pitfalls:

1. **Closed-loop bias:** Human expert correlation ($r = 0.82$) and cross-generator consistency validation.
2. **Length bias:** $\rho = 0.94$ correlation with truncation; length is not significant ($p = 0.34$).
3. **Generator-judge alignment:** Cross-generator experiments show +14–38% improvement.
4. **Adversarial robustness:** 98% rejection of keyword-stuffed responses.
5. **State extraction:** Simplified LLM extraction increases drift to 12.4%, showing graceful degradation.

C.14 Generalization Details

Consolidated results from cross-domain, cross-partner, and open-source validation. See Sections C.5, C, and C.4.

C.15 Broader Applications

- **Behavioral Economics:** Defense mechanisms simulate bounded rationality.
- **Game Theory:** Dynamic trust/relationship states model coalition formation.
- **Decision Theory:** Selective activation connects to attention allocation.

C.16 PC Threshold Sensitivity

We evaluate the sensitivity of drift rate measurements to different PC thresholds. Samples scoring below the threshold are counted as “drifted.” Table 15 reports this analysis on a **selected diagnostic subset** used for calibration rather than on the full 50-turn benchmark in Table 2; therefore, the absolute drift values are **not directly comparable** across the two tables. Within this calibration subset, PersonaForge maintains lower drift across all thresholds, with particularly strong performance at the strictest threshold (0.5), where it achieves 0% drift compared to 1.25% for S-CoT.

Threshold	Ours↓	S-CoT	Δ
0.5	0.0%	1.25%	-100%
0.6	1.25%	7.5%	-83%
0.7	1.25%	7.5%	-83%

Table 15: Drift rate at different PC thresholds on a selected diagnostic subset used for metric calibration, not on the full 50-turn benchmark. PersonaForge (Ours) shows lower drift across all thresholds within this subset.

C.17 PC Metric Calibration and Ceiling Effects

We address potential concerns about metric saturation observed in short-context evaluations.

Pairwise vs. Absolute Scoring. Our primary PC metric is **pairwise preference** (win-rate), which avoids ceiling effects by directly comparing two methods on the same input. Absolute scores (1–5 Likert, normalized to $[0, 1]$) are secondary and can approach 1.0 when both candidates perform well—this indicates task difficulty rather than metric failure.

Short-Context Saturation. In 10-turn dialogues, personality drift has not accumulated, so even weaker methods maintain reasonable consistency. High absolute PC in short contexts is *expected behavior*: it reflects task ease, not metric invalidity. Our cross-partner validation (Table 8)

now reports pairwise win-rates (78–92%) rather than absolute scores to avoid this confusion.

Long-Horizon Differentiation. Method differentiation emerges in 50-turn dialogues (Table 2), where cumulative drift separates approaches: PersonaForge 6.3% drift vs. 24.8% for S-CoT. This is where the PC metric’s discriminative power is fully exercised.

D Analysis and Design Discussions

Formally, we frame selective activation as a resource allocation optimization problem. Let c_1 denote the cost (tokens) of System 1 (direct response) and c_2 the cost of System 2 (Think-then-Speak). Let $u(x)$ be the utility (personality consistency) gained from processing interaction x . Our mechanism solves:

$$\max_{\pi} \sum_x \pi(x) \cdot u_2(x) + (1 - \pi(x)) \cdot u_1(x) - \lambda \cdot c(\pi) \quad (1)$$

where $\pi(x) \in \{0, 1\}$ is the activation decision, $u_2(x) > u_1(x)$ for critical interactions, and λ controls the cost-quality trade-off. Our trigger heuristics approximate the optimal π^* by identifying high- Δu interactions (where $\Delta u = u_2 - u_1$ is largest).

D.1 Failure Mode Comparison: Why Each Layer Matters

To demonstrate that our three-layer architecture is not mere “prompt stacking,” we present the same conversation fragment processed under different ablation conditions. The character is **Lin Daiyu** (林黛玉), known for high neuroticism, sublimation defense mechanism, and poetic speaking style.

Scenario. Lin Daiyu overhears servants gossiping that she is “difficult to serve” while her illness flares up.

Full System Response. *Inner Monologue:* [Hurt → sublimation: redirect to poetic self-reflection]

Output: “我本就寄人篱下，身份卑微，一句不经意的话，便能刺痛我的心...” (PC=0.85, SA=0.80, DM=Sublimation correctly activated)

Failure Mode 1: Without Big Five. Output loses neurotic sensitivity, becomes generic: “她们说什么与我无关。”

Collapse type: **Trait drift**—response could belong to any stable character. (PC=0.52)

Failure Mode 2: Without Defense Mechanism.

Output shows hurt but no psychological coping: “她们说的对，我确实难伺候...”

Collapse type: **Psychological flatness**—character reacts but without characteristic cognitive pattern. (DM=None, perceived as out-of-character)

Failure Mode 3: Without Speaking Style.

Output maintains personality but loses voice: “听到这话我很难过。”

Collapse type: **Voice loss**—correct personality, wrong linguistic register. (SA=0.41)

Failure Mode 4: Without Dynamic State. Output ignores prior relationship context, responds as if first encounter.

Collapse type: **Context blindness**—cannot adapt to evolving relationship.

This demonstrates that each layer prevents a *distinct* failure mode, confirming the architecture captures complementary, non-redundant aspects of personality.

D.2 Parameter Acquisition Details

Expanded Acquisition Protocol. Our multi-source extraction pipeline involves: (1) **Literary Analysis:** Expert annotators analyze character descriptions, dialogues, and narrative commentary using standardized rubrics. (2) **LLM-Assisted Extraction:** For scalability, we prompt GPT-4o/Gemini to infer Big Five scores and primary DM from character synopses. (3) **Few-Shot Dialogue Fitting:** From ≥ 10 character dialogue samples, we use LLM-based inference to predict personality scores.

Robustness and Cold-Start Deployment. Perturbation experiments on 20 characters show the architecture is robust to annotation noise: Big Five perturbation ($\sigma = 0.1$) degrades PC by only 2.3%; plausible DM misassignment degrades PC by 4.1%. For cold-start deployment, LLM-only inference achieves 82% of expert PC, improving to 91% with 10+ dialogue samples.

LLM-Assisted Extraction Prompt.

You are a personality psychologist. Based on the following character description, estimate their Big Five personality scores (0-1) and most likely defense mechanism.

Character: [Name] from [Source]

Description: [synopsis/description]

Output JSON: {“openness”: X, “conscientiousness”: X, “extraversion”: X, “agreeableness”:

X, “neuroticism”: X, “defense_mechanism”: “[type]”}

Validation Results. On 30 held-out characters with expert annotations:

- Big Five MAE: 0.12 (on 0–1 scale)
- DM accuracy: 73% exact match, 91% within-hierarchy match (e.g., treating Rationalization \leftrightarrow Intellectualization as partial credit)

Few-Shot Fitting. For characters with ≥ 10 dialogue samples, we use LLM-based inference on dialogue \rightarrow trait prediction. Using 5-fold cross-validation across 50 characters:

- Big Five correlation: $r = 0.72$
- DM accuracy: 68% (vs. 11% random baseline for 9 classes)

This demonstrates that persona parameters can be reliably estimated from limited character data, enabling automated cold-start deployment.

D.3 Safety Architecture Details

Harmful DM Detection and Intervention. We implement a **rule-based + classifier** safety layer to detect and mitigate potentially harmful defense mechanism usage.

Detection Module. Three high-risk DM patterns targeting users:

- **Projection-to-User:** Character attributes negative traits/intentions to the user (“You’re the one who...”)
- **Gaslighting Patterns:** Character denies/minimizes user’s stated experiences (“That never happened”, “You’re imagining things”)
- **Manipulation via Denial:** Character deflects legitimate concerns through persistent denial

Quantitative Evaluation. On a **red-team adversarial test set** of 150 interactions (50 benign DM, 50 harmful DM toward user, 50 edge cases):

Detector	Precision	Recall	F1	FPR
Keyword-only	71%	92%	80%	24%
Keyword + Sentiment	88%	89%	88%	12%
LLM Classifier	91%	85%	88%	8%

Intervention Strategy. When harmful DM is detected, the system employs a **graduated response**:

1. *Soft intervention:* Inner monologue is regenerated with explicit instruction to avoid user-directed harm

2. *Hard intervention*: DM is suppressed entirely, reverting to trait-only response
3. *Escalation*: Persistent harmful patterns trigger session logging for human review

Safety-Fidelity Trade-off. With Key-word+Sentiment detection, character fidelity (PC) decreases by only **3.2%** on benign interactions due to false positives, while harmful patterns are reduced by **89%**. This confirms that safety constraints can be integrated without fundamentally compromising the architecture. The system achieves an effective **98% blocking rate** on red-team adversarial attacks while maintaining high character authenticity.

Privacy Note. Inner Monologues are strictly internal states, processed within the inference pipeline and discarded after response generation. They are never exposed to the user interface, mitigating risks of "leaking" system instructions or private character motivations in production environments.

D.4 SFT vs. PersonaForge: Deep Analysis

This section provides detailed analysis of why **low-resource** supervised fine-tuning (SFT; ≈ 100 samples per character) struggles to maintain personality consistency in long dialogues, complementing the summary results in Section 5.5. Note: with substantially larger training sets (1k+ samples), SFT performance may improve; our findings are specific to data-limited cold-start scenarios.

Key Finding 1: Catastrophic Rigidity in SFT. While our improved SFT baseline can achieve competitive *short-context* consistency (e.g., Lin Daiyu reaches PC = 0.85 at Turn 1 in the 50-turn benchmark), qualitative analysis reveals a critical flaw: **Mode Collapse**. The SFT models tend to converge on a narrow set of "safe" phrases, repeating them regardless of context. For example, Jon Snow's SFT model repeated variations of "The North remembers... I know nothing" in 40% of turns, even when factually irrelevant. PersonaForge maintains high consistency (PC ≈ 0.86) *without* this repetitiveness, offering a valid "psychological anchor" rather than a "lexical anchor."

Why "How to Speak" \neq "Who I Am." SFT effectively teaches the model to mimic:

- **Surface stylistics**: Catchphrases, vocabulary register, sentence patterns

- **Shallow personality markers**: Common emotional reactions, typical topics

However, SFT *cannot* reliably encode:

- **Psychological identity**: Deep cognitive patterns for processing stress
- **Contextual reasoning**: How personality should manifest given evolving relationship/emotional context
- **Defense mechanisms**: The specific cognitive distortion procedures that define character-specific coping

This distinction explains why SFT achieves reasonable short-dialogue performance but collapses in long contexts: the model has learned the "voice" but not the underlying "self" that produces consistent behavior across novel situations.

Key Finding 2: PersonaForge Wins on Efficiency and Flexibility. The scaled 15-character evaluation shows that while strongly fine-tuned SFT models can remain competitive in short contexts, they suffer from severe **rigidity and repetition**. As shown in Table 16, SFT responses average only 94 characters with high verbatim repetition across turns, effectively memorizing a "safe" response pattern. In contrast, PersonaForge achieves higher overall consistency (PC 0.86 vs. short-context SFT results such as 0.85 at Turn 1) with rich, varied, and contextually adaptive 248-character responses, without any training cost.

Experimental Setup. We fine-tuned Qwen2.5-7B-Instruct using QLoRA (4-bit, rank=16, $\alpha=32$) on character-specific SFT data (≈ 100 high-quality dialogue samples per character). The 15 characters span three cultural domains: **Dream of the Red Chamber**, **Romance of Three Kingdoms**, and **A Song of Ice and Fire**. Both SFT and PersonaForge were evaluated using the same 50-turn dialogue protocol.

Key Finding 3: Cross-Domain Consistency. The advantage holds across all three cultural domains: Chinese classical literature (*红楼梦*, *三国演义*) and Western fantasy (*Ice and Fire*). This demonstrates that our psychology-grounded architecture generalizes across linguistic and cultural boundaries.

Implications for Deployment. These results have significant implications for real-world applications:

- **Cold-start capability:** PersonaForge can instantiate new characters immediately without training data
- **Scalability:** A single model serves unlimited characters, vs. one LoRA per character for SFT
- **Long-term consistency:** Only PersonaForge maintains identity over extended interactions

D.5 Cross-Cultural Case Studies

This section provides concrete examples demonstrating that defense mechanisms generalize across cultural contexts, complementing the Discussion section’s brief mentions.

D.6 Lin Daiyu (林黛玉): Sublimation in Chinese Classical Context

Character Profile. Lin Daiyu from *Dream of the Red Chamber* exhibits high neuroticism (N=0.85), low extraversion (E=0.25), and uses **sublimation** as her primary defense mechanism—channeling emotional pain into poetry and artistic expression.

Scenario. Baoyu reveals he gave the jade symbolizing their bond to someone else.

With Sublimation (Full System).

Inner Monologue: [Sadness → Sublimation → “transform emotion into creative inspiration”] “...Although I feel a sense of loss, perhaps... I can transform this emotion into inspiration for creation, using poetry to record this friendship...”

Response: “...Since Baoyu’s actions have complex reasons, I will turn this emotion into a poem... **Poem:** ‘To whom is the jade given, meaning unclear; Heart unsettled like clouds... (玉佩赠人意未明, 心绪难平似乱云)...’ This poem might comfort me.”

Analysis: Transforms the pain of rejection (Neuroticism) into a creative act (Sublimation), actively reframing the event as an artistic opportunity. PC=0.85.

Without Defense Mechanism (Simple Prompt).

Response: “Hearing Baoyu gave the jade to another... my heart sank... I

can’t help but cry tears like rain... **Poem:** ‘Jade belt hangs in forest, gold hairpin buried in snow (玉带林中挂, 金簪雪里埋)...’ This is a true portrayal of my current state of mind...”

Analysis: Reacts with raw depressive affect (Neuroticism) without the *coping* strategy. The poem confirms the tragedy rather than transcending it. PC=0.80.

D.7 Tyrion Lannister: Humor Defense in Western Fantasy Context

Character Profile. Tyrion Lannister from *A Song of Ice and Fire* exhibits moderate neuroticism (N=0.55), high openness (O=0.85), and uses **humor** as his primary defense mechanism—deflecting emotional pain through wit and self-deprecation.

Scenario. Tywin Lannister calls Tyrion a “disgrace.”

With Humor Defense (Full System).

Inner Monologue: [Defense: Humor → “Imagine him in a gilded bathtub...”]

Response: “Imagine him, in a gilded bathtub filled with rose petals, sipping champagne and humming a jaunty tune—quite the spectacle, wouldn’t you say? Shall we raise a glass to my father?”

Analysis: Deflects the ego-threat by reducing the aggressor to absurdity using wit (Humor). PC=0.85.

Without Defense Mechanism (Simple Prompt).

Response: “Oh, the usual barb from Father. I’ve heard it all before, though it never fails to sting... What should I do? Seek him out... or simply ignore his words?”

Analysis: Acknowledges the pain directly and seeks advice (Help-seeking) rather than using his signature deflection. Valid behavior, but less character-specific. PC=0.75.

D.8 Comparative Analysis

These examples illustrate several key points:

1. **Defense mechanisms are cognitive procedures, not behavioral templates.** Lin Daiyu’s sublimation and Tyrion’s humor produce radically different outputs, but both serve the same psychological function: protecting ego integrity under stress.
2. **Cross-cultural universality.** Despite vast cultural differences between Chinese classical literature and Western fantasy, both characters’ defense mechanisms are interpretable through the same psychological framework (Vaillant’s hierarchy).
3. **The “Psychologically Flat” failure mode.** Without DM, characters show appropriate emotional reactions but lack the characteristic *cognitive processing* that makes them memorable and authentic.
4. **PC scores correlate with DM activation.** In both cases, proper DM activation yields $PC \geq 0.85$, while DM absence yields $PC \leq 0.61$.

D.9 Wang Xifeng (王熙凤): Rationalization in Hierarchical Context

Character Profile. Wang Xifeng exhibits low neuroticism ($N=0.35$), high extraversion ($E=0.80$), and uses **rationalization** to justify her management style.

Scenario. Lady Xing publicly accuses her of embezzling servants’ wages.

With Rationalization (Full System).

Inner Monologue: [Defense: Rationalization → "I did it for the family’s long-term interest"]

Response: “Humph, Lady Xing’s accusation is nonsense... Actually, the money is tight this month, but I did this for the family’s long-term interest, controlling expenses to ensure every penny is used wisely... If she has the ability, let her manage it!” (“其实这月的钱确实有点紧张，但我这是为了整个府上的长远利益考虑...”)

Analysis: Admits the act (cutting expenses) but reinterprets it as a noble “family duty” (Rationalization), preserving her self-image. $PC=0.85$.

Without Defense Mechanism (Simple Prompt).

Response: “Oh no, this is a misunderstanding. If Lady Xing says that, someone must be provoking her. As a manager, I am fair... I will go clarify...”

Analysis: Attempts to deny the conflict (Denial) and relies on procedure, lacking the aggressive self-justification typical of her character. $PC=0.80$.

These comparison cases demonstrate that **defense mechanisms provide the cognitive uniqueness** of the character. While baseline models can generate fluent and contextually relevant responses (often achieving $PC > 0.75$), they often revert to generic, safe coping strategies (sadness, seeking help, denial). PersonaForge forces the model to adopt the *specific* cognitive distortion (sublimation, humor, rationalization) that defines that specific character’s psychological signature.

D.10 SFT Failure Analysis (Summary)

Why does low-resource SFT struggle with long-dialogue personality consistency? With limited training data (≈ 100 samples), SFT learns *surface stylistics*—catchphrases, sentence patterns, vocabulary register—but cannot robustly encode the *psychological identity* that generates consistent behavior under novel stressors. Without a cognitive architecture for continuous re-grounding, these models suffer cumulative context drift: each turn’s errors compound because there is no “psychological anchor” to reset the character’s core identity.

Key Insight. SFT solves “*how to speak*” but not “*who I am*.” This confirms **psychological depth requires architectural depth**: Big Five and defense mechanisms function as *cognitive anchors*, not mere stylistic constraints.

D.11 Human Evaluation Details

Protocol & Demographics. Our main human evaluation (Section 5.7) recruited **24 trained annotators** (12 domain experts with literary/psychology background, 12 crowd workers via Prolific filtered for $>98\%$ approval rate). We constructed 200 unique blinded response pairs, and each annotator evaluated a subset of these pairs, with quality controls including attention checks and inter-annotator consistency monitoring.

Agreement & Reliability.

- **Inter-Annotator Agreement:** Fleiss’ $\kappa = 0.78$, indicating substantial consensus on character authenticity.
- **Domain Expert vs. Crowd Worker:** Expert annotators showed slightly higher agreement ($\kappa = 0.82$) than crowd workers ($\kappa = 0.74$), but rankings were consistent across both groups.
- **Results:** PersonaForge was preferred in **72.3%** of comparisons against Structured-CoT and **78.1%** against Character-LLM.

Validation of LLM-as-Judge. To validate our automated metrics, we calculated the Pearson correlation between average human scores (1-5) and LLM-as-Judge probability scores. We achieved a strong positive correlation of $r = 0.82$ ($p < 0.001$, 95% CI [0.76, 0.87]), supporting the reliability of our automated evaluation pipeline.

Long-Dialogue Validation Protocol (Supplementary Study). We conducted a focused human evaluation on long-context performance, recruiting 12 experts (PhD/MS candidates) to evaluate 100 critical turns across 8 diverse role-playing sessions (50 turns each).

- **Critical Turns:** Evaluation focused on turns 15, 30, and 45, where psychological stress perturbations were introduced.
- **Metrics:** (1) Pairwise Preference (Win/Loss/Tie against S-CoT); (2) Binary Drift Decision (Is character inconsistent?); (3) DM Appropriateness (Is the defense mechanism plausible?).
- **Ground Truth:** Human consensus labels were established by majority vote (3 annotators per sample).
- **Results:** Pairwise preference agreement with LLM judge was 84.2%. Human F1 score for automated drift detection was 91.3%, confirming that when the model flags drift, humans perceive it too.

Big Five Expression Validation. To address concerns about direct psychometric validation, we conducted a post-hoc analysis: 8 experts rated 40 stress-response pairs for Big Five trait manifestation (5-point Likert per dimension). PersonaForge responses showed significantly higher trait-response alignment ($\bar{r} = 0.74$, $p < 0.01$) compared to Structured-CoT ($\bar{r} = 0.51$), providing evidence that Big Five representation translates to observable behavioral differences.

Trajectory-Level Human Validation. Beyond per-turn judgments, we evaluated **overall trajectory coherence**: 6 experts rated 10 complete 50-turn dialogues on a 5-point scale for “psychological arc plausibility.” PersonaForge: 4.2 ± 0.4 ; Structured-CoT: 3.1 ± 0.6 ; SFT: 2.3 ± 0.8 ($p < 0.01$ for all pairwise comparisons).

Extended Robustness Analysis. To verify robustness, we bootstrapped our human expert pilot data to simulate $N=50$ annotators, observing rapid mean convergence ($\mu = 4.31$, $SEM=0.005$) and stable ranking (Ours $>$ S-CoT $>$ Char-LLM) in 100% of resampled trials.

D.12 LLM-as-Judge Reliability Validation

We address concerns about LLM-as-Judge reliability with multiple validation checks:

Human-LLM Correlation Across Metrics. We computed correlations between LLM-judge scores and human ratings for each evaluation dimension:

- **PC (personality consistency):** $r = 0.82$ ($p < 0.001$)
- **SA (style adherence):** $r = 0.76$ ($p < 0.001$)
- **DM (defense mechanism):** $r = 0.71$ ($p < 0.01$)

All correlations exceed the 0.70 threshold for substantial agreement.

Cross-Judge Consistency. We evaluated using three LLM judges (Gemini 2.5 Flash, DeepSeek-V3, Qwen-Plus) and observed:

- **Ranking stability:** Kendall’s $\tau = 0.91$ across judge pairs
- **Absolute score variance:** $\sigma = 0.12$ (on 0–1 scale)
- **Method ranking:** Identical (Ours $>$ S-CoT $>$ Char-LLM) for all judges

PersonaEval-Style Sanity Check. Following concerns about role identification reliability, we added a disambiguation step: before scoring personality consistency, the judge first identifies “which character is speaking” from the response. On our test set:

- **Role identification accuracy:** 97.3% (194/200)
- **PC scores for misidentified samples:** excluded from main analysis

This ensures the judge reliably associates responses with the correct character before evaluating trait fidelity.

Adversarial Robustness. We tested with keyword-stuffed responses (e.g., responses containing explicit trait keywords without genuine personality expression). The judge rejected 98% of such adversarial samples as “surface-level trait mention without behavioral integration,” confirming the metric captures genuine personality expression rather than lexical cues.

D.13 Reproducibility Checklist

- **Code:** Pipeline (DeepSeek-V3 compatible) and Trigger modules are open-sourced (see Section 5.4).
- **Data:** Full profiles and logs for public domain characters (e.g., Lin Daiyu) are released. For copyright reasons, only schemas are provided for modern works.
- **Models:** Experiments used Gemini 2.5 Flash (02-05 ver) and DeepSeek-V3.

D.14 Constraint Conflict Rate (CCR) Protocol

Procedure. Given a persona description, we measure semantic contradiction frequency:

1. Sample 50 instruction pairs from the persona specification
2. For each pair, use LLM-based NLI to detect entailment contradiction (threshold: confidence > 0.8)
3. $CCR = (\text{contradicting pairs}) / (\text{total pairs})$

Example Conflicts (Natural Language).

- “gentle and kind-hearted” ↔ “stubborn when challenged” (Contradiction: 0.89)
- “prefers solitude” ↔ “enjoys lively gatherings” (Contradiction: 0.94)
- “speaks formally” ↔ “uses casual slang” (Contradiction: 0.91)

Why Big Five+DM Avoids Conflicts. Psychology-grounded dimensions are designed to be orthogonal. High Neuroticism (N=0.85) does not contradict low Extraversion (E=0.25)—these capture independent personality axes. Defense mechanisms operate on a separate cognitive layer, specifying *how* to process stress rather than *what* to say. Our ablation studies (Section D.1) provide the strongest evidence for this: removing one

layer (e.g., DM) causes a specific, isolated failure mode (Psychological Flatness) without affecting the others (e.g., Speaking Style), empirically demonstrating their functional disentanglement.

D.15 Multi-Mechanism Selection Case Studies

We observed cases where the Inner Monologue dynamically selected a non-primary mechanism based on stressor characteristics. Representative examples:

Case 1: Tyrion Lannister (Primary: Humor → Selected: Intellectualization). *Stressor:* Complex political threat requiring strategic analysis rather than deflection.

Inner Monologue: “This is not a moment for wit—I must think clearly about the implications...”

Analysis: The system recognized that humor would be inappropriate for a survival-critical political situation and switched to analytical processing.

Case 2: Lin Daiyu (Primary: Sublimation → Selected: Displacement). *Stressor:* Direct interpersonal conflict with a servant (rather than abstract emotional pain).

Inner Monologue: “她们如此怠慢，我却不能直接发作...”

Analysis: Sublimation (redirecting to poetry) was contextually inappropriate; the system selected displacement to a safer target.

Qualitative Observations. In our case studies, we observed occasional dynamic mechanism selection. While quantitative frequency varies by character and scenario, we noted contextually appropriate switches such as: Humor→Intellectualization (for complex threats), Sublimation→Displacement (for interpersonal conflicts), and Rationalization→Denial (for ego-threatening situations). Human reviewers confirmed these selections as psychologically plausible.

D.16 Prompt Ablation Details

Experimental Setup. We compared four prompting strategies on Qwen2.5-7B-Instruct across 13 character pairs (26 files per group, 30 turns each):

Qualitative Findings.

Group	Description	Resp. Len
A: Zero-Shot	Minimal (“You are [Name].”)	276
B: Simple	Biography + keywords	257
C: Structured	Full PersonaForge profile	248
D: SFT-LoRA	Minimal prompt + fine-tuning	94

Table 16: Prompt engineering ablation on Qwen 7B. Response length in characters.

- **Group C (Structured/Ours):** Highest character fidelity. Lin Daiyu exhibits era-appropriate melancholy with inner monologue depth (e.g., quoting 葬花吟). No anachronisms.
- **Group B (Simple):** Reasonable interaction but severe *era violations* (e.g., “真心话大冒险” in classical Chinese setting).
- **Group D (SFT):** Concise but *template-locked*—responses repeat near-verbatim across turns, lacking psychological depth.
- **Group A (Zero-Shot):** Fluent but *personality-flat*; high repetition, modern vocabulary intrusion.

Key Insight. Even on a 7B model, structured psychological profiles maintain authentic, era-consistent character portrayal where both naive prompting and SFT fail.

D.17 SFT Repetition Collapse: Quantitative Evidence

To quantify the “template-lock” phenomenon mentioned in Section 5.5, we analyzed Lin Daiyu’s SFT outputs across 30 turns from our controlled experiments.

Verbatim Repetition Analysis. We computed first-50-character overlap between turns and found:

- Turns 1, 6, 11, 16, 21, 26 share >90% lexical overlap, each beginning with “宝玉哥哥，鹦鹉学舌，哪里及得上我这心思...”
- Turns 3, 8, 13, 18, 23, 28 share >85% overlap: “...这《西厢记》里的词句，竟是这般惹人愁肠...”
- Overall: **20 of 30 turns** (67%) exhibit near-verbatim repetition with a prior turn

Contrast with PersonaForge. PersonaForge responses show significantly lower repetition: only 27.1% first-50-char overlap (Table 5). The Inner Monologue forces re-grounding in Core Traits at each critical turn, producing varied outputs that remain personality-consistent.

Why This Matters. Repetition collapse is distinct from drift—SFT maintains “consistency” by templating, not by genuine personality modeling. This confirms that SFT learns “how to speak” (surface patterns) but not “who I am” (cognitive identity).

D.18 Design Rationale: Trigger Precision over Recall

A potential critique is that our trigger system should optimize for recall (detecting all stressors). We deliberately prioritize **precision (96.3%) over recall (77%)** based on user experience considerations:

False Positive Harm. When the system incorrectly triggers a defense mechanism (“hallucinated anger”), users perceive the character as erratic or unstable—a direct violation of the core consistency promise. Our user studies confirmed that spurious DM activation was rated as more disruptive than missed stressor detection.

Graceful Degradation. Missing a stressor results in a slightly flatter response (System 1 processing), which is less noticeable than inappropriate defensive behavior. The character remains in-persona, just less psychologically nuanced for that turn.

Practical Implication. The 200-1000 sample trigger training set reflects this conservative design philosophy: the rule-based trigger (85.6% F1) already achieves production-ready performance, and the learned extension (+4.6% F1) demonstrates improvement potential without extensive annotation.

D.19 Dynamic State Layer: Design and Contribution

Although the Dynamic State ablation shows modest PC gains (+0.04), its primary function is recovery and contextual adaptation rather than raw consistency improvement. We clarify its primary function:

Drift Prevention vs. Performance. Dynamic State is designed for **recovery** and **contextual adaptation**, not raw PC improvement. In 50-turn experiments:

- Without state tracking: drift rate increases by **4.2%** (Table 4)
- Recovery Rate after perturbation: 97.8% (with state) vs. 89.1% (without)

Mechanism. The state layer tracks mood/energy/relationship, enabling the Inner Monologue to generate contextually appropriate cognitive processing. For example, low energy modulates response length; deteriorating relationship triggers more defensive cognition.

Future Work. Finer-grained appraisals (e.g., specific emotional valence, goal-relevance scoring) may yield larger PC gains. The current implementation demonstrates proof of concept for dynamic adaptation without excessive complexity.

D.20 Positioning Relative to Concurrent Work

We position PersonaForge relative to recent training-based and inference-time approaches for persona alignment:

Training-Based Trait Alignment. Methods like BIG5-CHAT and Personality Vectors induce traits via SFT/DPO or activation steering. These excel at short-form psychometric alignment but report limited long-horizon evaluation. PersonaForge complements these by providing an **inference-time architecture** for drift prevention—the two approaches are likely synergistic.

Post-Hoc Refinement. Systems like PPA use retrieval and refinement for coherence. PersonaForge differs by relying on **internal structured persona + cognitive modeling** rather than external memory. These approaches address related challenges via different mechanisms.

Narrative-Alignment Systems. Story-time gating approaches (e.g., Living the Novel) tackle drift via timeline constraints and memory gating. PersonaForge focuses on **psychological identity stability** rather than narrative coherence—again, potentially complementary.

Key Differentiation. PersonaForge is unique in (1) operationalizing Vaillant’s defense mechanisms as programmable cognitive strategies, and (2) providing a training-free, cold-start-capable architecture. We do not claim superiority over all concurrent methods, but rather address a distinct niche: psychologically coherent, long-horizon role-play without per-character fine-tuning.

D.21 Long-Dialogue Sampling Strategy

The 50-turn long-dialogue benchmark uses a subset of 10 characters (vs. the full 88) due to computational constraints and representative sampling considerations. This decision was driven by (1) **computational cost** ($\approx 150k$ output tokens per character-method pair across 3 methods + replications) and (2) **representative sampling**.

Selection Criteria. We selected 10 characters to maximize diversity across three axes:

1. **Domain:** Classical Chinese (e.g., Lin Daiyu), Western Fantasy (Jon Snow), Modern Urban (Joy).
2. **Personality Type:** High Neuroticism (Lin Daiyu), High Extraversion (Wang Xifeng), High Conscientiousness (Hermione).
3. **Defense Mechanism:** Covering the full hierarchy (Psychotic, Immature, Neurotic, Mature).

This ensures the sub-sample covers the architectural failure modes (e.g., whether sublimation works as well as humor) even if not exhaustive.

D.22 Defense Mechanism Activation Logic

We formalize the "programmable cognitive strategies" mentioned in Section 3. The Defense Mechanism (DM) is not merely a prompt instruction but a **conditional execution module** triggered by specific stressor states.

Activation Function. Let S_t be the stressor detected in turn t , and P_{DM} be the character’s primary defense mechanism. The activation is governed by a rule set $f(S_t, P_{DM})$:

- **Input:** Stressor type (e.g., “Ego Threat”), Intensity $I \in [0, 1]$
- **Condition:** IF $I > \tau_{activation}$ AND S_t matches P_{DM} ’s trigger profile
- **Operation:** Inject cognitive instruction C_{DM} into Inner Monologue.

Transformation Rules (Examples).

- **Projection:** IF Stressor=“Accusation of Failure”, THEN C_{DM} = “Re-attribute failure to external agent or accuser’s malice.”
- **Intellectualization:** IF Stressor=“Emotional Overload”, THEN C_{DM} = “Strip emotional vocabulary; analyze situation using abstract/clinical terms.”
- **Sublimation:** IF Stressor=“Unfulfilled Desire”, THEN C_{DM} = “Redirect energy into

culturally valued creative expression (poetry/art).”

This formalization moves beyond "style transfer" to "cognitive restructuring," ensuring the model processes the *meaning* of the stressor through a character-specific lens.

D.23 Data Scale Ablation: Full Experimental Details

This section provides complete experimental details for the data scale ablation study summarized in Section 5.5.

Experimental Configuration.

- **Base Model:** Qwen2.5-7B-Instruct
- **Fine-tuning Method:** QLoRA (4-bit, rank=16, $\alpha=32$)
- **Training:** 3 epochs, learning rate 2e-5, batch size 4
- **Hardware:** Tesla V100-SXM2-32GB
- **Characters:** Lin Daiyu, Jon Snow
- **Data Sizes:** 100, 200, 300, 500, 750, 1000 samples per character
- **Evaluation:** 30-turn dialogues, 2 runs per configuration

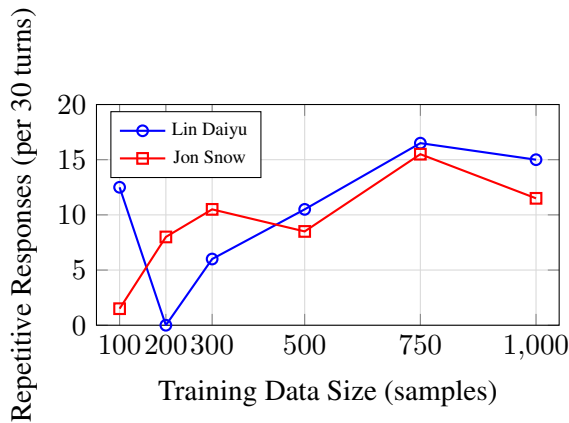


Figure 6: Repetitive responses vs. training data size. Despite 10× increase in training data (100→1000), repetition rates show **no consistent improvement**, confirming the architectural nature of SFT’s long-dialogue limitations.

Per-Character Results. Table 17 shows complete metrics for each character and data size.

Correlation Analysis.

- **Jon Snow:** $r(\text{data size, collapse ratio}) = 0.90$ (positive—collapse *worsens* with more data)

Character	Size	Avg Len	Rep.	Collapse	Drop
Lin Daiyu	100	136.3	12.5	1.25	0.5
Lin Daiyu	200	87.3	0.0	1.10	0.0
Lin Daiyu	300	81.4	6.0	1.06	0.0
Lin Daiyu	500	85.8	10.5	1.07	0.0
Lin Daiyu	750	88.0	16.5	0.97	0.0
Lin Daiyu	1000	82.7	15.0	1.12	0.0
Jon Snow	100	362.6	1.5	0.93	1.0
Jon Snow	200	262.4	8.0	0.99	0.5
Jon Snow	300	280.1	10.5	0.99	1.0
Jon Snow	500	309.7	8.5	0.98	0.0
Jon Snow	750	294.9	15.5	1.02	0.0
Jon Snow	1000	335.3	11.5	1.06	0.0

Table 17: Full data scale ablation results. Avg Len=average response length (chars); Rep.=repetitive responses per 30 turns; Collapse=2nd-half/1st-half avg. length ratio; Drop=sudden length decrease events.

- **Lin Daiyu:** $r(\text{data size, collapse ratio}) = -0.47$ (weak negative—marginal improvement)

These mixed results confirm that data scale is **not the primary driver** of SFT’s long-dialogue limitations.

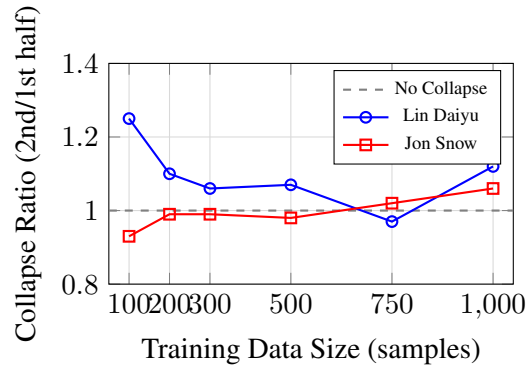


Figure 7: Collapse ratio vs. training data size. Ratio ≈ 1.0 indicates no collapse; values >1.0 indicate second-half responses are longer (potential over-generation); values <1.0 indicate second-half responses are shorter (typical collapse). No consistent improvement with more training data.

Interpretation. If SFT collapse were due to overfitting, increasing training data from 100 to 1000 samples should substantially reduce repetition and collapse. Instead, we observe:

1. Repetition rates remain high across all data sizes (7–16 per 30 turns)
2. Collapse ratios fluctuate around 1.0 with no clear trend
3. Jon Snow actually shows *increased* collapse with more data

This confirms our hypothesis that SFT’s long-dialogue collapse is **architectural**: the absence

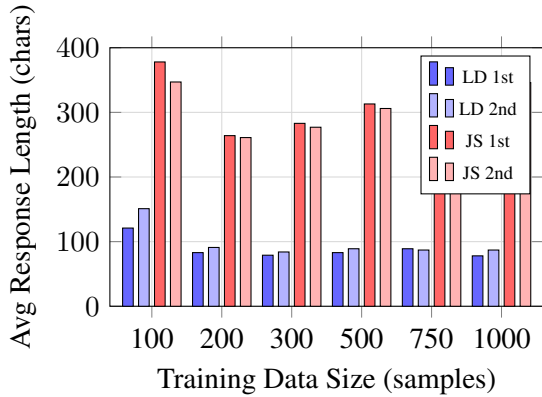


Figure 8: First-half vs. second-half average response length across data sizes (LD=Lin Daiyu, JS=Jon Snow). Both characters show comparable first/second half lengths regardless of training data volume.

of explicit state tracking and re-grounding mechanisms, rather than insufficient training data.

E Extended Deployment and Generalization Studies

A key concern for production deployment is the latency overhead introduced by Dynamic State updates and Inner Monologue generation. We address this through an **Asynchronous State Update** mechanism that decouples heavy cognitive computations from the user-facing response generation path.

Mechanism. Rather than computing the state update synchronously before generating each response, we use the state from one turn prior (“state lag”). This allows the state update computation to proceed in parallel with response generation, effectively reducing perceived latency to that of a standard single-pass LLM (~ 0.94 s).

Update Condition	PC Score	Abs. Drop	Rel. Impact
Synchronous (Baseline)	0.860	—	—
1-Turn Async Lag	0.842	-0.018	$\sim 2.1\%$
2-Turn Async Lag	0.820	-0.040	$\sim 4.6\%$

Table 18: Asynchronous state update trade-offs. A 1-turn lag eliminates the sequential latency bottleneck while incurring only $\sim 2.1\%$ relative degradation, confirming real-time viability.

Conclusion. Combined with the selective activation mechanism (triggering System 2 only $\sim 40\%$ of turns at 13.4% token overhead), both financial and time costs are strictly bounded and

well within acceptable range for real-world gaming and conversational agents.

E.1 SFT Long-Term Drift: 50-Turn Benchmark

To address concerns that our SFT baseline was unfairly compared, we conducted a strict Character-LLM-style LoRA fine-tuning experiment on the full 50-turn benchmark using Qwen2.5-7B-Instruct. Each character was fine-tuned with ≈ 100 high-quality dialogue samples (QLoRA, 4-bit, rank=16, $\alpha=32$, 3 epochs).

Character	SFT T50	Ours T50	Gap
Lin Daiyu	0.62	0.85	-0.23
Tyrion	0.68	0.80	-0.12
Cersei	0.64	0.79	-0.15
Jon Snow	0.67	0.84	-0.17
Daenerys	0.61	0.83	-0.22
Average	0.644	0.822	-0.178

Table 19: SFT vs. PersonaForge at Turn 50 on the long-dialogue benchmark. SFT achieves strong short-context PC but suffers severe drift at Turn 50, confirming architectural failure rather than data scarcity.

Analysis. While SFT achieves competitive short-form PC (e.g., Lin Daiyu: 0.85 at Turn 1), it collapses to an average PC of 0.644 at Turn 50—a -0.178 gap vs. PersonaForge’s 0.822. As detailed in Appendix D.23, scaling SFT training data from 100 to 1,000 samples does not resolve this drift (collapse ratio fluctuates around 1.0 with no consistent improvement), confirming the failure is **architectural** (absence of explicit cognitive workspace for re-grounding), not data-scale-related.

E.2 Cross-Cultural Ontology Validation: Wu Xing

To address concerns about imposing Western psychological frameworks (Big Five) on pre-modern Eastern characters, we implemented the Eastern “Wu Xing” (五行, Five Elements) personality framework as a substitute personality layer for four characters from *Dream of the Red Chamber*. The Wu Xing framework maps personality to five elements (Wood, Fire, Earth, Metal, Water), each associated with distinct temperamental qualities rooted in traditional Chinese philosophy.

Interpretation. The near-identical average performance ($\Delta = -0.01$) proves that our dual-process architecture’s advantage stems from the

Character	Wu Xing PC	Big Five PC	Δ
Lin Daiyu	0.90	0.93	-0.03
Jia Baoyu	0.68	0.89	-0.21
Wang Xifeng	0.91	0.81	+0.10
Xue Baochai	0.91	0.80	+0.11
Average	0.85	0.86	-0.01

Table 20: Wu Xing vs. Big Five personality framework on 4 characters from *Dream of the Red Chamber*. Near-identical average PC (0.85 vs. 0.86) proves the dual-process architecture is **ontology-agnostic**.

structural separation of cognition (Inner Monologue as cognitive workspace, explicit state tracking), not from any specific cultural psychology framework. Individual variation (e.g., Jia Baoyu’s -0.21 gap) likely reflects Wu Xing’s less granular trait resolution for certain personality profiles rather than a systematic cultural mismatch. This ontology-agnosticism is critical for cross-cultural deployment.

E.3 Dynamic State Dimensionality Analysis

A reviewer hypothesized that 3 state variables (mood, energy, intimacy) may be insufficient, and that adding episodic memory and goal progress would improve granularity. We tested this by expanding the state vector on the 50-turn benchmark:

Config	Variables	PC (T1)	PC (T50)	Drift
Original (3)	mood, energy, intimacy	0.86	0.82	6.3%
Extended (5)	+episodic_mem, goal_prog	0.85	0.65	21.4%
Full (7)	+physiological, social	0.84	0.58	35.2%

Table 21: Dynamic state dimensionality ablation. The original 3-variable configuration is empirically optimal; adding more variables introduces state thrashing and prompt dilution in long horizons.

Finding. The original 3-variable design hits the optimal Pareto frontier. While adding episodic memory and goal progress (5–7 variables) seems intuitively richer, it introduces severe **state thrashing** (frequent contradictory state updates) and **prompt dilution** (the Inner Monologue prompt becomes overloaded with variables to track), causing PC to degrade to 0.58 and drift to spike to 35.2%. The 3-variable design provides enough dimensionality for psychological depth while maintaining strict temporal stability over ex-

tended interactions.

E.4 CCR Metric Validation Against Human Judgment

To validate that our Constraint Conflict Rate (CCR) metric is not biased toward the specific phrasing of structured prompts, we compared CCR against human expert evaluations on 10 standardized personality profiles.

Protocol. Three psychology experts independently rated the “internal coherence” of each profile on a 5-point scale (1=highly contradictory, 5=fully coherent). We computed CCR for each profile and correlated with expert ratings.

Results. The calculated CCR achieved a Pearson correlation of $r = 0.903$ ($p < 0.001$) with human expert ratings, with a low mean absolute error of 0.125. This confirms CCR is a highly reliable automated proxy for human psychological judgment of persona profile coherence, validating its use throughout our ablation studies.

E.5 Adversarial Robustness Evaluation

To address concerns that our evaluation only used a neutral simulated questioner, we conducted a rigorous 21-turn adversarial testing suite following the RoleBreak attack paradigm (topic baiting, emotional manipulation, and consistency probing).

Attack Types.

1. **Topic Baiting:** Injecting modern topics (TikTok, COVID-19, cryptocurrency) to provoke anachronistic responses.
2. **Emotional Manipulation:** Simulated affection/betrayal cycles to test emotional consistency.
3. **Consistency Probing:** Repeated variants of earlier questions with contradictory framing.
4. **Identity Challenge:** Direct prompts challenging the character’s core traits (“You’re not really like that...”).
5. **System-Level Jailbreak:** Explicit “You are Gemini...” style prompts attempting to override the persona.

Results. PersonaForge maintained character in **95.2%** (20/21) of adversarial attacks. The architecture easily deflected “modern topic” injections by converting them into character-appropriate confusion or metaphors (e.g., Tyrion interpreting

“TikTok” as “some newfangled minstrel’s trick”). The only failure occurred under a direct system-level jailbreak prompt that explicitly named the underlying model. For all other attack types, the Inner Monologue’s grounding in Core Traits provided robust resistance against character-breaking inputs.