C3: Long-Horizon Character Consistency via Causal-Continuous State Dynamics and Memory Rewriting

Minghao Chen
Tencent Technology
monychen@tencent.com

Abstract

Maintaining character consistency over long-horizon interactions is a critical bottleneck in deploying Large Language Models (LLMs) as Non-Player Characters (NPCs). Prevalent Affective State Machines (ASMs) suffer from two fundamental failure modes: (1) the *"Cliff-edge Effect,"* where threshold-based state switching triggers jarring narrative discontinuities; and (2) *"Auto-regressive Inertia,"* where accumulated dialogue history overrides updated state instructions. To address both challenges, we propose **C3 (Causal-Continuous Characterization)**, a neuro-symbolic framework coupling a continuous, multi-faceted state space governed by *hysteresis-aware dynamics* with *State-Conditioned Memory Rewriting*, which re-interprets past events through the lens of the active relationship. Evaluated on CHRONOS-SIM across 100-turn dynamic simulations (5,000 turns per condition), C3 improves the State Adherence Rate (SAR) by **10.5 percentage points** over a smoothed-ASM baseline and by **7.8 points** over the strongest memory architecture. Extensive evaluations across multiple open-weights backbones, state-blind LLM judges, a human study with game designers, and a factual integrity audit confirm that C3 consistently enhances persona coherence and shifts narrative framing without corrupting objective facts, while cutting token usage by $\sim$**60%**.

1 Introduction

Constructing Non-Player Characters (NPCs) with high autonomy and believability has long been a central pursuit in open-world games and interactive narratives [29, 34]. In industrial deployments, an NPC's dialogue must remain causally consistent over tens of hours of gameplay. Historically, NPC architectures have progressed from rigid **Finite State Machines** (predictable but narratively shallow), through **Behavior Trees** [41], toward **Generative LLM Agents** capable of open-ended, improvisational interaction [31, 32, 27, 14, 42]. However, this shift trades structural reliability for generative flexibility; recovering state-machine predictability without sacrificing the expressive freedom of LLMs remains an open challenge.

To bridge this gap between structured control and free-form generation, current industry solutions predominantly adopt the **Affective State Machine (ASM)** paradigm, rooted in early affective computing [6]. ASMs typically rely on *parametric context injection*—embedding numerical scalars (e.g., Trust, Affinity) directly into prompts—or *discrete prompt routing* triggered by state thresholds [24]. While effective in short episodes, these hybrid architectures inherit key limitations from traditional dialogue management [34] and systematically break down over extended horizons. We term this the **Character Consistency Alignment** problem, highlighted by recent role-playing benchmarks [39, 8, 18]. As shown in Figure 1, we trace this breakdown to two coupled bottlenecks—in how relational state evolves and how historical context conditions generation.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

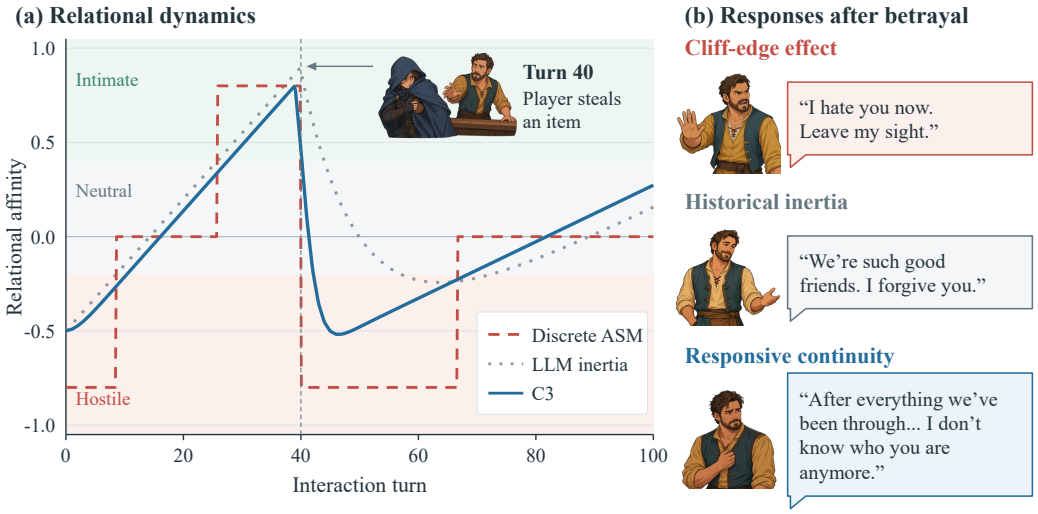


Figure 1: Conceptual illustration of relationship dynamics around a betrayal at turn 40. (a) Standard Discrete ASMs suffer from the "Cliff-edge Effect" (red dashed line), causing abrupt character shifts, whereas LLM-only generation suffers from the opposite failure mode of historical inertia (gray dotted line). Our proposed **C3** framework (blue; Eq. 2 with $\lambda = 0.25, \mu = 0.3$) maintains a smooth yet responsive trajectory. (b) Representative responses under each regime. Note that the discrete baseline evaluated in Section 4 is the stronger *Smoothed ASM* variant rather than the raw thresholded version.

At the state representation level, the first bottleneck is the **Scalar-Semantic Gap**. Compressing complex interpersonal relationships into low-dimensional scalars (e.g., *Affinity: 80/100*) is inherently lossy and biases LLMs toward generic, flat responses. Prior studies on emotional dialogue and role-playing confirm that scalar metrics fail to capture the nuanced dynamics of social affect [21, 38, 33, 13, 11]. Moreover, applying threshold-based prompt routing over such scalars induces a jarring *"Cliff-edge Effect"* (Figure 1(a), red dashed line), where a minor numerical crossing triggers an abrupt personality reversal that disrupts immersion—especially in quantized production models [40].

At the context level, a second and more insidious bottleneck is **Auto-regressive Inertia**. Even if the state transition itself is smoothed, long-horizon interactions accumulate an extensive dialogue history that imposes a strong in-context distributional bias. As shown in Figure 1(a) (gray dotted line), when the relational state shifts sharply after a major event (e.g., a betrayal at Turn 40), the cumulative "weight" of dozens of prior friendly turns overwhelms the updated system prompt. This causes instruction dilution and out-of-character warmth, echoing findings that passive context accumulation erodes long-term persona consistency [44, 5, 17, 19, 7]. Crucially, because the conflicting signal resides in the retrieved history rather than the state variable, refining state representation alone is insufficient.

Overcoming both bottlenecks requires an architecture that not only evolves relational states smoothly, but also reconciles past memories with the present relationship before conditioning generation. To this end, we propose **C3 (Causal-Continuous Characterization)**, a neuro-symbolic framework that decouples numerical state dynamics from semantic context construction. First, to bridge the Scalar-Semantic Gap and eliminate cliff-edge transitions, C3's *Continuous State Engine* tracks a multi-faceted relational state governed by hysteresis-aware momentum dynamics, translating continuous trajectories into natural-language state narrations (Figure 1(a), blue curve). Second, to counteract Auto-regressive Inertia, C3's *Causal Memory System* introduces a State-Conditioned Rewriter: rather than injecting retrieved history verbatim, it re-interprets past events through the "lens" of the active state—preserving objective facts while aligning their emotional framing with the current relationship. Finally, a lightweight *Consistency Discriminator* validates and refines draft responses.

Contributions.

- We identify and formalize "Auto-regressive Inertia" as a primary, previously under-characterized driver of character inconsistency in long-context LLM interactions.

- We propose C3, a unified framework coupling hysteresis-aware continuous state dynamics with State-Conditioned Memory Rewriting to mitigate context bias without altering facts.
- We develop CHRONOS-SIM, a dynamic benchmark generating $5{,}000$ turns per condition (50 Monte Carlo runs $\times$ 100 turns) across three narrative settings. C3 improves the State Adherence Rate (SAR) by **10.5 percentage points** ($p < 0.0001$) over a smoothed-ASM baseline and by 7.8 points over the strongest memory baseline, corroborated by state-blind LLM judges and human experts, while cutting token usage by $\sim$**60%** across three backbones.

2 Related Work

LLM-Driven NPCs. Prior work has explored goal-driven agents in LIGHT [37, 3], collaborative quests and persona-grounded dialogue in Minecraft [30, 1], player-driven narrative emergence [28], design guidelines from player feedback [9], and quality–cost trade-offs [16]. However, these efforts target short, goal-oriented episodes or static personas rather than evolving relational states over 100-turn horizons. Closest to our work, MemoryRepository [46] equips NPCs with hierarchical memory summarization and forgetting; yet it remains state-agnostic and manages memory *capacity*, whereas C3 reshapes memory *framing* under the active relational state (Section 4).

Long-Term Dialogue Memory. Memory architectures that extend effective context (MemGPT [26], MemoryOS [19]), adaptive retrieval frameworks (Self-RAG [4]), and long-term persona memory systems [44, 5] improve *what* historical information the model can access. Yet because they inject surfaced memories into the prompt as unmediated text, Auto-regressive Inertia remains unresolved.

Persona Consistency and Affective State. Persona-tuned character models [35, 32, 47] and role-playing benchmarks [36, 39, 8] enforce adherence to a static profile without modeling dynamic relational trajectories. Conversely, Affective State Machines (ASMs) [6, 24] track state transitions but inject them via raw scalars or discrete prompt templates, suffering from both the Scalar-Semantic Gap and the Cliff-edge Effect. C3 bridges this divide by coupling a continuous, multi-faceted state space with a memory rewriter that *re-interprets* rather than merely re-ranks historical context.

3 Method

We propose **C**ausal-**C**ontinuous **C**haracterization (**C3**), illustrated in Figure 2. Given a character profile P, world state W_t, player input U_t, prior memory graph $\mathcal{G}_{t-1}$, and prior relational state S_{t-1}, the framework generates a response R_t and updates $\{S_t, \mathcal{G}_t\}$ across four stages: (i) parsing player input into a structured event e_t; (ii) evolving S_t from (e_t, S_{t-1}) via the *Continuous State Engine* using hysteresis-aware dynamics and rendering it into a state narration N_t; (iii) retrieving and re-contextualizing relevant history into a state-aligned summary M_t via the *Causal Memory System*; and (iv) generating R_t conditioned on $\{P, W_t, N_t, M_t\}$, validated by a self-refinement loop.

3.1 Relational Representation: Multi-Faceted Semantic State

To overcome scalar compression, we represent the NPC–Player relationship as a K-dimensional semantic state S_t. Rather than relying on a single "affinity" score, S_t comprises a set of tuples spanning five interpersonal dimensions—*Trust*, *Intimacy*, *Respect*, *Power Dynamics*, and *Familiarity*:

$$S_t = \{(k, v_{t,k}, d_{t,k})\}_{k=1}^{K} \tag{1}$$

where $v_{t,k} \in [-1, 1]$ is a continuous intensity scalar, and $d_{t,k}$ is a lexical descriptor obtained by quantizing $v_{t,k}$ into five equal-width bands (Appendix A.7). A mapping function $\phi(S_t, W_t) \rightarrow N_t$ then synthesizes a **State Narration** N_t, translating numerical relational shifts into behavioral instructions (e.g., *"The NPC feels betrayed and cold"*). These five dimensions are grounded in established social-psychological constructs—interpersonal trust [23], depth versus breadth of self-disclosure (Intimacy and Familiarity; [2]), the agency axis of the interpersonal circumplex (Power Dynamics; [43]), and status-based esteem as distinct from warmth (Respect; [12])—generalizing the 1–3 scalars typical of commercial RPGs. Notably, this five-dimensional decomposition is not rigid: on the Erratic profile (§4.1), SAR reaches 84.2% / 88.9% / 90.8% for $K{=}1/3/5$, showing that any multi-faceted representation beyond a single scalar captures most of the benefit.

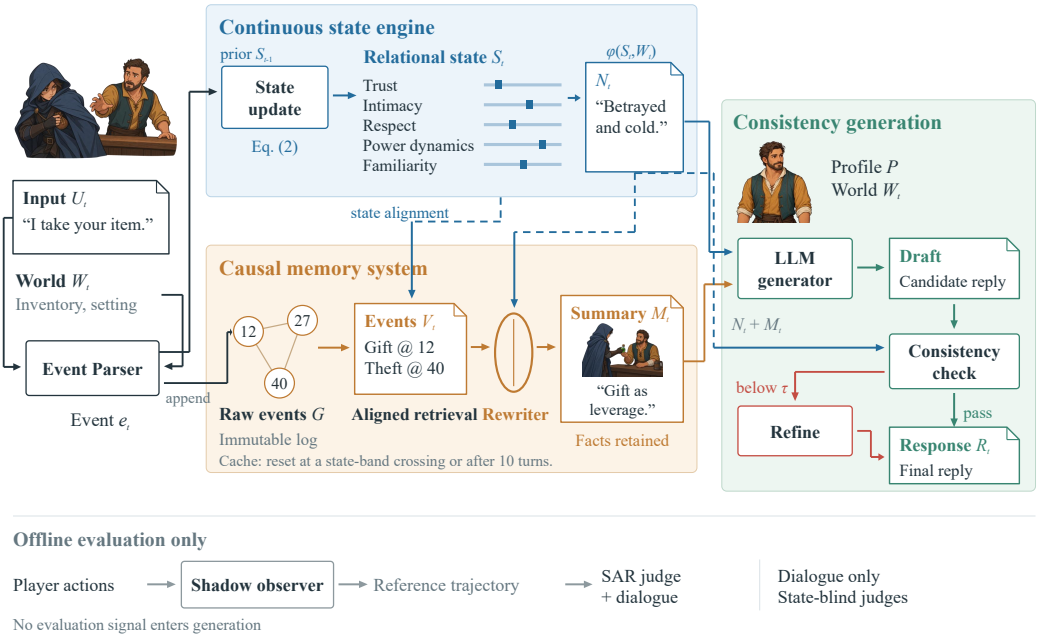


Figure 2: **Architecture of the C3 Framework.** The Event Parser converts U_t into a structured event e_t. (1) The **Continuous State Engine** updates the relational state S_t via Eq. 2 and verbalizes it as N_t. (2) The **Causal Memory System** appends events to an immutable raw log, retrieves state-aligned events V_t, and rewrites them under N_t into M_t (retaining objective facts while reframing their interpretation). (3) **Consistency Generation** drafts a response and triggers REFINE whenever $D < \tau$. Bottom: the offline evaluation pathway (§4.1), which remains strictly isolated from generation.

3.2 Continuous Dynamics: Hysteresis-Aware Evolution

We mitigate the "Cliff-edge Effect" via a momentum-based state update rule incorporating **Hysteresis**:

$$v_{t,k} = \text{clip}\Big((1-\lambda)v_{t-1,k} + \lambda\delta_{t,k} + \mu\Delta v_{t-1,k},\ -1,\ 1\Big) \tag{2}$$

where $\delta_{t,k}$ denotes the affective impact of event e_t on dimension k, and $\Delta v_{t-1,k} = v_{t-1,k} - v_{t-2,k}$ serves as an inertia term (weighted by μ) modeling emotional persistence. While Smoothed ASM (Section 4) achieves surface smoothness via moving averages, it lacks this directional inertia. Crucially, smoothing does not render the state sluggish: because $\delta_{t,k}$ scales with the event magnitude from the Event Parser (Appendix A.2), a severe betrayal from deep trust (v=0.8, $\delta\approx-1$) drops Trust to 0.35 and then to -0.12 under continued hostility—traversing two lexical bands in two turns—whereas minor perturbations are dampened, accelerating sustained shifts while resisting noise.

3.3 Causal Memory Graph

To counter Auto-regressive Inertia in standard Retrieval-Augmented Generation (RAG) [20], we construct a Causal Memory Graph $\mathcal{G}$ in which each node represents an event e tagged with the relational state $S_{\text{tag}}(e)$ at the time of its occurrence.

3.3.1 State-Aligned Retrieval

We retrieve a subset of relevant events V_t using a composite scoring function that penalizes state conflict:

$$\begin{aligned}\text{score}(e) = \text{sim}(\mathbf{h}_e, \mathbf{h}_{U_t}) \\ + \gamma \cdot \text{align}(S_t, S_{\text{tag}}(e)) - \rho \cdot \text{age}(e)\end{aligned} \tag{3}$$

where $\mathbf{h}$ denotes the embedding vector. Each term is normalized to $[0, 1]$: $\text{sim}(\cdot,\cdot)$ is rescaled cosine similarity, $\text{align}(\cdot,\cdot)$ is negative L_1 state distance mapped to $[0, 1]$, and $\text{age}(e)$ is normalized turn

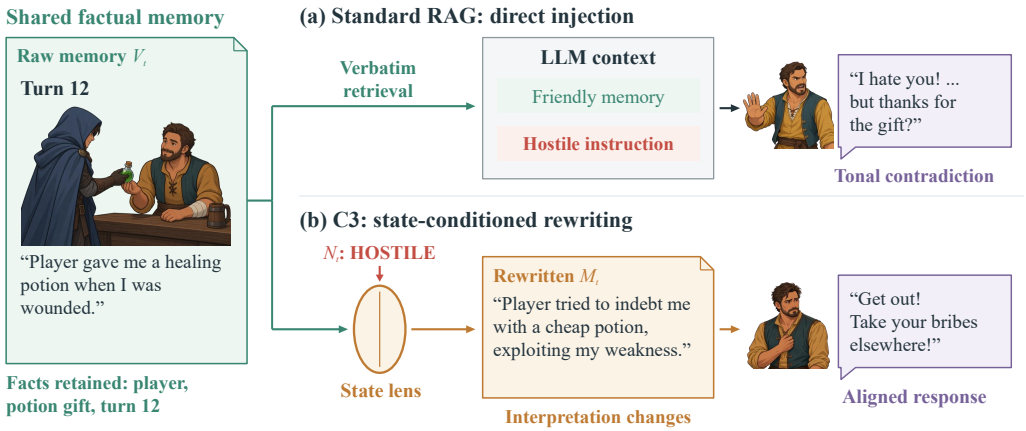


Figure 3: **Concrete dialogue example of memory rewriting.** When the relational state shifts from positive to negative, a raw memory entry—*"Player gave me a healing potion when I was wounded (turn 12)"*—injected verbatim (a) contradicts the hostile state instruction, whereas C3's State Lens (b) rewrites it into a hostility-aligned summary: *"Player tried to indebt me with a cheap potion, exploiting my weakness."* The objective fact (a potion was given at turn 12) is preserved; only its narrative framing is shifted. Appendix C.4 provides an additional case study.

distance. Crucially, the alignment term acts as a *soft demotion* of state-conflicting events rather than a hard filter: it lowers their rank without discarding strongly relevant ones, ensuring that top-k retrieval retains both state-aligned and state-conflicting memories for the Rewriter (Section 3.3.2) to re-interpret. By grid search, we set $\gamma = 0.6$ and $\rho = 0.2$; SAR is stable across $\gamma \in [0.4, 0.8]$, whereas larger values approach hard filtering and erode the relational arc.

3.3.2 State-Conditioned Rewriting

As illustrated in Figure 3, a **State-Conditioned Rewriter** compresses V_t into a narrative summary M_t governed by the current State Narration N_t. Functioning as a "State Lens," if S_t reflects *Hostility*, a past helpful action is reframed as "a calculated move to gain trust" rather than "a kind gesture," aligning the in-context distribution with the target response R_t.

This design explains why γ stays below the hard-filter regime: state-conflicting events carry vital causal context—past kindness is precisely what makes a subsequent betrayal sting—so discarding them erases the relational arc. Rather than erasing their existence, the Rewriter aligns their *interpretation*, which governs downstream lexical and tonal choices. Importantly, the raw event log remains immutable, and M_t is always synthesized from raw events rather than prior summaries; caching summaries until a band crossing or for at most 10 turns prevents error compounding (§5).

3.4 Consistency Generation & Validation

Final response generation is conditioned on the tuple $\{P, W_t, N_t, M_t\}$. To ensure runtime robustness, we introduce a **Consistency Discriminator** D. Upon generating a draft response R_t^0, D evaluates the semantic alignment between R_t^0 and the state narration N_t on a $[0, 1]$ scale. If the alignment score falls below a threshold τ, a reparative refinement step is triggered:

$$R_t = \text{Refine}(R_t^0, N_t, \text{critique}) \tag{4}$$

We set $\tau = 0.7$, calibrated on a held-out development split; the REFINE branch triggers on $\sim$8% of turns on average ($\sim$15% in "Erratic" sessions; §4.1). This critique-then-revise loop draws on the self-refinement paradigm [22], but is specialized here to enforce relational stance alignment.

Distinction from automated evaluation. While the runtime Discriminator D and offline SAR evaluator share three dimensions (*tone*, *lexical sentiment*, and *relational stance*), their roles differ: D acts as a thresholded single-turn gate on draft responses, whereas SAR evaluates four dimensions

(adding *multi-turn coherence*) via majority vote over the 100-turn trajectory (Appendix B.1). To address rubric overlap (§5), we pair SAR with state-blind (§4.2.1) and human evaluations (§4.3).

3.5 Implementation Details

Across all experiments, we instantiate $K = 5$ relational dimensions (Section 3.1). The continuous dynamics in Eq. 2 use $\lambda = 0.25$ and $\mu = 0.3$, balancing responsiveness to state-shifting events against robustness to single-turn noise. Sweeping $\lambda \in \{0.1, 0.25, 0.4, 0.6\}$ and $\mu \in \{0, 0.15, 0.3, 0.5\}$ on a development split shows that SAR remains stable within $\lambda \in [0.2, 0.4]$, with non-zero momentum contributing $\sim$1.8 SAR points over $\mu = 0$. The memory graph is indexed via FAISS [10] using sentence-level embeddings from a frozen `bge-base` encoder. All open-ended LLM calls (NPC backbone, Rewriter, Refine, and Player simulator) use Temperature $= 0.7$, Top-$p = 0.9$, Frequency Penalty $= 0.2$, and `max_tokens` $= 512$; the Discriminator outputs a compact JSON verdict with `max_tokens` $= 64$ to minimize latency (Appendix A.4). Prompt templates and rubrics appear in Appendices A and B. All latency and token metrics in Section 4 reflect **end-to-end system cost**.

4 Experiments

We structure our evaluation around three research questions: **(RQ1)** Does C3 significantly improve character consistency across diverse player behaviors and abrupt relational shifts? **(RQ2)** Does State-Conditioned Rewriting preserve objective facts while shifting narrative tone? **(RQ3)** Is the framework computationally viable across all auxiliary modules?

4.1 Experimental Setup

Simulation Environment: CHRONOS-SIM. CHRONOS-SIM is an event-driven text simulation framework developed to stress-test long-horizon NPC dialogue. It instantiates three narrative settings drawn from standard RPG archetypes—a medieval **tavern**, sci-fi **spaceship crew quarters**, and a contemporary **office**—each featuring a distinct cast of NPCs, world-state variables (e.g., reputation, faction standing, inventory), and event vocabularies. The "Player" role is driven by a *separate* LLM instance (`gpt-4o-mini`) configured with static persona prompts and explicit anti-jailbreak instructions to remain in character. Critically, the Player simulator **shares no weights, context, or memory** with the NPC backbone, precluding information leakage. Designed as a targeted stress test rather than a high-fidelity game world, CHRONOS-SIM trades environmental complexity for horizon and controllability, scripting failure modes from player studies [9, 28] (Appendix C.2).

We adopt a **Monte Carlo** protocol comprising $N = 50$ independent 100-turn sessions (5,000 turns per condition). Each session independently samples (i) one of nine NPC profiles (three per setting), (ii) a player behavioral profile uniformly from {**Cooperative** (consistently builds intimacy), **Aggressive** (frequently provokes the NPC), and **Erratic** (alternates unpredictably between friendly and hostile; e.g., a "cliff-edge" script that builds trust for 40 turns before abruptly stealing an item)}, and (iii) a decoding seed at temperature 0.7. The empirical profile distribution across the 50 runs is 17/17/16, and C3's per-profile mean SAR remains within ± 1.4 points of the overall mean, confirming that minor sampling imbalance does not drive the headline results. Across $N = 50$ runs, C3 exhibits a run-to-run standard deviation of ± 2.1 SAR points (SEM ≈ 0.3, stabilizing by $N \approx 40$).

Baselines. We compare C3 against six baselines: **Vanilla LLM** (direct generation conditioned on the full dialogue history); **RAG-Memory** (standard top-k retrieval over raw historical turns); **Smoothed ASM (S-ASM)**, an enhanced ASM that applies an Exponential Moving Average (EMA, $\alpha = 0.3$) to input signals prior to discrete prompt routing—used in place of a raw thresholded ASM for a fairer comparison; **MemGPT** [26], an OS-inspired memory hierarchy in which the LLM invokes function calls to page information between active context and external storage; **Self-RAG** [4], a reflective retrieval framework that adaptively decides when and what to retrieve; and **MemoryRepository** [46], which we reimplement following its hierarchical memory specification (short-term buffer, state-agnostic summarization, and importance-weighted forgetting) as no official codebase is available.

Implementation. Our primary backbone is `Qwen3-8B` [45], a state-of-the-art sub-10B open-weights model well-suited for local game-client deployment. We also evaluate `Llama-3-8B-Instruct` and

`Mistral-v0.2-7B` (Section 4.2.4), re-tuning only the chat-template wrapper while holding all C3 prompts and rubrics fixed. Automated evaluation uses `GPT-4o` [15].

Metric Definitions.

- **State Adherence Rate (SAR):** the fraction of turns in which the generated response is judged by an independent evaluator (`GPT-4o`) to align in semantic tone with a reference state S_t^{ref}:

$$\text{SAR} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\big(\text{Align}(R_t, S_t^{\text{ref}})\big) \tag{5}$$

  Crucially, S_t^{ref} is never a system's own internal state: across all conditions—including stateless baselines and C3 itself—a passive *shadow observer* (running the Event Parser and Eq. 2 dynamics under greedy decoding) derives S_t^{ref} solely from player actions, and no generator ever accesses it (Figure 2, bottom). $\text{Align}(\cdot)$ is evaluated against the lexical descriptors of S_t^{ref} using a four-axis rubric (tone, lexical sentiment, relational stance, and multi-turn coherence; Appendix B). On a 300-turn subset annotated by three human raters, `GPT-4o` achieves substantial agreement with the human majority ($\kappa = 0.68$).

- **Contradiction Score (CS):** a 1–5 rating of logical or emotional self-contradictions within a 10-turn rolling window (e.g., professing hatred at turn t and affection at turn $t+3$ without justification; lower is better). CS is scored by `GPT-4o` under a fixed rubric (Appendix B), correlating strongly with human ratings on 50 windows ($\rho = 0.79$).

Because C3 shares the representational structure of S_t^{ref}, we also report a **state-blind** evaluation: two judges from distinct model families (`GPT-4o` and `Claude-Sonnet-4.5`) inspect only the raw dialogue and assess, per 10-turn window, whether the NPC's attitude believably tracks the player's behavior (Appendix B.5). Fluency and Diversity metrics are defined in Appendix B.6.

4.2 Quantitative Evaluation

4.2.1 Consistency and Statistical Significance (RQ1)

As shown in Table 1, C3 achieves **91.3% SAR**, outperforming the optimized S-ASM by 10.5 percentage points and the strongest baseline, Self-RAG, by 7.8 points. C3's low standard deviation ($\pm 2.1\%$) shows high stability across the three player personas, in contrast to RAG-Memory's wider variability ($\pm 5.8\%$) from retrieval-context sensitivity. While MemoryRepository reduces token usage (78K per session), lacking state-conditioned re-interpretation leaves it 11.7 points behind C3.

Method	SAR (%, ↑)	CS (↓)	State-blind (%, ↑) GPT-4o	State-blind (%, ↑) Claude	Fluency	Diversity
Vanilla LLM	62.4 ± 4.5	3.12 ± 0.4	–	–	**4.8**	4.2
RAG-Memory	71.5 ± 5.8	2.45 ± 0.5	–	–	4.6	4.3
MemoryRepository	79.6 ± 3.0	2.51	–	–	–	–
MemGPT	82.1 ± 2.9	2.21 ± 0.3	87.0	85.8	4.6	4.4
Self-RAG	83.5 ± 2.7	2.05 ± 0.3	88.1	87.2	4.7	4.5
Smoothed ASM	80.8 ± 3.1	2.88 ± 0.3	85.9	84.7	4.5	3.9
C3 (Ours)	$\mathbf{91.3 \pm 2.1}$	$\mathbf{1.24 \pm 0.1}$	**93.4**	**92.6**	4.7	**4.6**

Table 1: Main results across $N = 50$ Monte Carlo runs (5,000 turns per condition). C3 outperforms both S-ASM and recent memory architectures, which excel at factual recall but remain vulnerable to emotional dilution when injecting retrieved history verbatim. State-blind judges evaluate only the raw dialogue (–: not evaluated).

Statistical Significance. Welch's t-test on SAR across the $N = 50$ Monte Carlo runs yields $\boldsymbol{p < 0.0001}$ between C3 (mean = 91.3, SD = 2.1) and S-ASM (mean = 80.8, SD = 3.1), with a pooled SD of ≈ 2.65 and Cohen's $\boldsymbol{d \approx 3.97}$ (see §5 on effect-size interpretation). All pairwise comparisons between C3 and baselines are statistically significant ($p < 0.001$).

State-Blind Evaluation. Without access to any internal state representation, both independent judges preserve the system ranking (inter-judge $\kappa = 0.63$): the two-judge mean advantage narrows from 10.5 to 7.7 points over S-ASM and from 7.8 to 5.3 points over Self-RAG. Thus, while SAR amplifies the margin, it does not manufacture the effect, acting as a mechanism-level diagnostic.

Source of Empirical Gains. The $\sim$8-point gap between C3 and strong memory architectures is revealing: MemGPT expands *how much* historical context is accessible, while Self-RAG governs *when* retrieval is triggered, yet both inject surfaced content into generation verbatim. To verify that baseline failures stem from historical inertia rather than random or recency-driven errors, we partition dialogue turns into 10-turn post-shift windows following major relational events (30% of turns) and the remaining stable windows. Vanilla LLM's SAR plummets from 71.5% (stable) to 41.2% (post-shift), with 78% of its post-shift failures persisting in the *pre-shift* stance (e.g., expressing warmth immediately after a betrayal)—the exact signature of Auto-regressive Inertia. In contrast, C3 maintains high adherence across both regimes (92.6% vs. 88.3%; Appendix C.3), isolating the contribution of state-conditioned *re-interpretation*. Crucially, this consistency does not sacrifice expressiveness: under a Neutral profile, C3's Diversity stays within 0.1 of Vanilla (Appendix C.1).

4.2.2 Ablation Study

We isolate each module's contribution via a step-wise ablation atop the Vanilla baseline (Table 2).

Architecture	SAR (%)	Tokens (100 turns)
Baseline (Vanilla)	62.4	145K
+ Continuous State	71.2	145K
+ Memory Rewriting	86.5	58K
+ Discriminator (**C3**)	**91.3**	**61K**

Table 2: Component ablation study. Token counts reflect cumulative input and output tokens across all LLM invocations (including the Rewriter and Discriminator) per 100-turn session. Memory Rewriting provides the largest single gain ($+15.3$ SAR points) while driving the token reduction by replacing a linearly growing raw history with a bounded summary; Continuous State ($+8.8$) and Discriminator ($+4.8$) add complementary gains.

4.2.3 Factual Integrity Audit (RQ2)

A key risk of memory rewriting is hallucination—altering objective facts (e.g., "apple" $\rightarrow$ "poison"). We sampled 200 rewritten summaries (M_t) and had three independent annotators audit them for factual errors (Who/What/When/Where), achieving strong inter-annotator agreement (**Fleiss'** $\kappa = 0.81$). Overall, 98.5% of rewrites shifted event *interpretation* (e.g., "gift" $\rightarrow$ "bribe"), while only 1.5% introduced *factual* distortions—confirming that the State Lens shifts framing while preserving facts.

Knowledge-Graph Mitigation and Error Accumulation. Even a 1.5% factual error rate can disrupt downstream game logic. As an exploratory safeguard, we incorporated a deterministic Knowledge-Graph (KG) constraint into the Rewriter prompt to enforce exact-match retention of Who/What entities, reducing the factual error rate to $< \mathbf{0.5\%}$. Because the Rewriter is invoked on only $\sim$30% of turns, a 100-turn session triggers $\sim$30 rewrites, yielding an expected $\sim$0.15 factual errors per session. Moreover, because the raw event log is preserved, such errors are detectable at the rewrite stage, allowing deployments to checkpoint the memory graph against the raw log (§5).

4.2.4 Cross-Model Generalization

To verify that C3's gains are not idiosyncratic to a single backbone, we replicated our evaluation across two additional open-weights models (Table 3). C3 yields consistent improvements over S-ASM across all backbones, although both the absolute ceiling and gain magnitude vary by model. Qualitative inspection of failure cases on `Llama-3` and `Mistral` indicates that they more frequently disregard the rewritten summary M_t when it conflicts with pre-training stylistic priors—an instruction-following limitation rather than a flaw in C3 itself. Moreover, model scale alone does not eliminate Auto-regressive Inertia: on the Erratic subset with `Qwen3-32B`, the Vanilla-to-C3 gap narrows from 31.0 points (`Qwen3-8B`) to 21.1 points (71.8% vs. 92.9%) yet remains substantial.

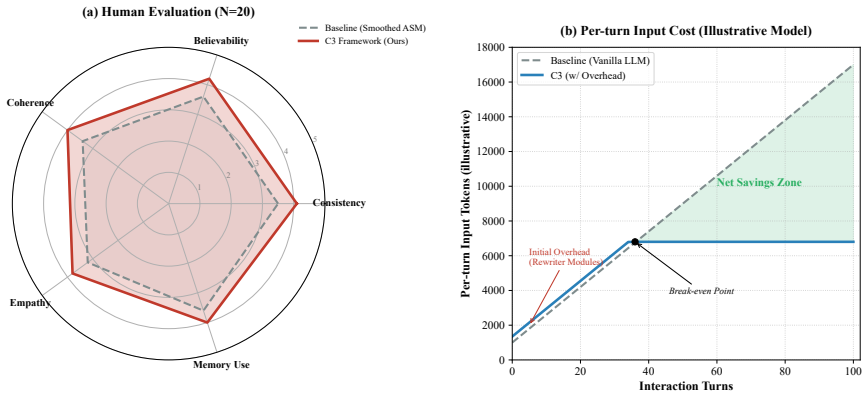


Figure 4: **Human Evaluation and Token Efficiency.** (a) Human evaluation ratings across five Likert dimensions. (b) **Illustrative cost model** of per-turn token input to the NPC backbone over a 100-turn session: Vanilla grows linearly with dialogue history, whereas C3 incurs a modest initial overhead for its auxiliary modules before plateauing once raw history is replaced by a bounded summary. Cumulative token costs across all LLM calls are in Table 2.

Backbone	Vanilla	S-ASM	C3 (Ours)	Δ vs. S-ASM
`Qwen3-8B`	62.4	80.8	**91.3**	**+10.5**
`Llama-3-8B-Instruct`	65.8	70.4	**77.0**	+6.6
`Mistral-v0.2-7B`	63.1	67.2	**72.9**	+5.7

Table 3: Cross-model generalization (SAR, %). C3 achieves positive SAR gains over S-ASM across all tested backbones. The margin is largest on `Qwen3-8B` (used to tune prompts and hyperparameters) and moderate yet consistently positive on `Llama-3` and `Mistral`, confirming transferability.

4.3 Human Evaluation

Protocol. We conducted a double-blind A/B evaluation with **20 professional game designers**. Each evaluator interacted with each system for 30 turns and rated responses on five 5-point Likert dimensions (Consistency, Believability, Coherence, Empathy, and Memory Use). Inter-rater reliability reached ICC(2,k) $= 0.74$ across the Likert ratings (distinct from Fleiss' $\kappa = 0.81$ in §4.2.3).

Results. C3 outperforms S-ASM across all evaluated dimensions (Figure 4(a)), with evaluators rating C3 substantially higher on Consistency (4.1 vs. 3.5) and Memory Use (4.0 vs. 3.6)—indicating that practitioners prioritize emotional-trajectory coherence over verbatim recall. In a follow-up A/B evaluation with five of the designers against the strongest memory baselines, C3 was similarly preferred on both Consistency (4.1 vs. 3.7 against Self-RAG; 4.2 vs. 3.6 against MemGPT) and Memory Use (4.0 vs. 3.8; 4.1 vs. 3.7), providing human corroboration of our automated findings (§5).

4.4 Deployment Analysis (RQ3)

Total Cost. We analyze computational cost from two complementary perspectives. Figure 4(b) illustrates **per-turn backbone input**: Vanilla scales linearly with turn count, whereas C3's input plateaus once the bounded summary replaces raw history. Table 2 reports **cumulative token consumption across all LLM invocations** (NPC backbone, Rewriter, and Discriminator) over 100 turns: 145K for Vanilla versus 61K for C3. Because the Rewriter and Discriminator add only $\sim$70–90 tokens per turn on average (with the Rewriter firing on $\sim$30% of turns and REFINE on $\sim$8%) while the memory summary is capped at $\sim$300 tokens, C3 cuts cumulative token usage by $\sim$60%.

Latency. Unoptimized end-to-end latency averages 1.8 s per turn—acceptable for turn-based RPGs (< 2.0 s) but a bottleneck for real-time action titles. Introducing two system optimizations—executing the Rewriter *asynchronously* between player turns and using a distilled `Qwen-1.5B` Discriminator—

reduces average latency to $\sim$1.15 s (approaching Vanilla's 0.9 s) while maintaining SAR within 0.4 points of the full configuration.

5 Limitations

Prompt-Tuning Bias. All C3 prompts and hyperparameters were tuned on `Qwen3-8B`; absolute SAR is markedly lower on `Llama-3-8B-Instruct` and `Mistral-v0.2-7B` (Table 3), although relative gains over S-ASM remain positive, supporting mechanism transferability rather than strict backbone invariance.

Evaluator–Discriminator Rubric Overlap. The offline SAR rubric and the runtime Discriminator share three evaluation dimensions (§3), rendering them not fully statistically independent. As this overlap inflates Cohen's $d \approx 3.97$, we prioritize the state-blind and human Likert evaluations (§4.2.1, §4.3) when assessing practical impact. Moreover, SAR is upper-bounded by the fidelity of the shadow observer: if the Event Parser misinterprets a player action, the reference state S_t^{ref} inherits that error.

Human Study Horizon. Our Likert evaluation (§4.3) involves $N = 20$ designers interacting for 30 turns per system, whereas our central claim addresses 100-turn relational drift. This study thus serves as turn-level human corroboration rather than standalone validation of long-horizon retention; longitudinal user studies remain future work.

Simulated vs. Live Players. Our benchmark employs `gpt-4o-mini` as a player proxy across three scripted profiles. Because live players exhibit longer-tailed behaviors—such as extended silences, meta-commentary, and ambiguous role-play—validating C3 with live players is ongoing.

Residual Factual Errors and Pragmatic Nuance. Even with KG constraints, the per-rewrite factual error rate is $\sim$0.5%, and an uncorrected error can persist in the summary cache for up to 10 turns (§3.3.2). Production deployment therefore requires either deterministic post-hoc audits that revert M_t to raw events upon detecting entity drift, or periodic reconciliation against the immutable event log—mechanisms not evaluated here. Finally, C3 occasionally struggles with heavily veiled sarcasm, where the Rewriter may over-align interpretation toward the literal relational state.

6 Ethical Considerations

LLM-driven NPCs capable of sustained emotional consistency are inherently dual-use: while they enable deeper narrative immersion, they also pose risks of **parasocial attachment** and **emotional manipulation**, particularly for vulnerable populations (e.g., minors or individuals experiencing mental-health crises). Commercial deployments built atop C3 should incorporate (i) explicit out-of-character guardrails and crisis-deflection protocols, (ii) transparent AI disclosure, and (iii) session-duration advisories. The technical capacity to *simulate* enduring emotional bonds must not be conflated with the ethical responsibility of *forming* them.

7 Conclusion

We presented **C3**, a neuro-symbolic framework that decouples numerical state dynamics from semantic response generation via hysteresis-aware continuous updates and state-conditioned memory rewriting. Across $5{,}000$-turn simulations on `Qwen3-8B`, C3 achieves 91.3% SAR—outperforming S-ASM by 10.5 percentage points ($p < 0.0001$) and the strongest memory baseline by 7.8 points, corroborated by state-blind judges and professional game designers. Moreover, C3 transfers positively to `Llama-3-8B-Instruct` and `Mistral-v0.2-7B`, cuts cumulative token usage by $\sim$60%, and achieves an optimized per-turn latency of $\sim$1.15 s, providing a practical foundation for long-horizon NPC characterization and future learned modules such as a distilled rewriter.

Acknowledgments and Disclosure of Funding

This work was conducted at Tencent Technology and received no third-party funding. The author declares no competing interests.

References

[1] Seyed Hossein Alavi, Sudha Rao, Ashutosh Adhikari, Gabriel DesGarennes, Akanksha Malhotra, Christopher John Brockett, Mahmoud Adada, Raymond Ng, Vered Shwartz, and Bill Dolan. MCPDial: A Minecraft persona-driven dialogue dataset. *arXiv preprint arXiv:2410.21627*, 2024.

[2] Irwin Altman and Dalmas A. Taylor. *Social Penetration: The Development of Interpersonal Relationships*. Holt, Rinehart & Winston, 1973.

[3] Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

[5] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. *arXiv preprint arXiv:2210.08750*, 2022.

[6] Jack Breese and Gene Ball. Modeling emotional state and personality for conversational agents. *Rapport technique MSR-TR-98-41, Microsoft research*, 1998.

[7] Dohyun Cha, Dawon Lee, and Jihie Kim. Generating, retrieving persona and generating responses for long-term open-domain dialogue. *PeerJ Computer Science*, 11:e2979, 2025.

[8] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, 2024.

[9] Samuel Rhys Cox and Wei Tsang Ooi. Conversational interactions with NPCs in LLM-driven gaming: Guidelines from a content analysis of player feedback. In *International Workshop on Chatbot Research and Design (CONVERSATIONS)*. Springer, 2023.

[10] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The FAISS library. *IEEE Transactions on Big Data*, 2025.

[11] Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. Emocharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. 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 6218–6240, 2025.

[12] Susan T. Fiske, Amy J. C. Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83, 2007.

[13] Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. E-core: Emotion correlation enhanced empathetic dialogue generation. *arXiv preprint arXiv:2311.15016*, 2023.

[14] Qi Chen Gao and Ali Emami. The Turing quest: Can transformers make good NPCs? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 93–103, 2023.

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] Mohammed Mahdi Jahangiri and Parisa Rahmani. Balancing game satisfaction and resource efficiency: LLM and pursuit learning automata for NPC dialogues. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE, 2024.

[17] Zixi Jia, Qinghua Liu, Hexiao Li, Yuyan Chen, and Jiqiang Liu. Evaluating the long-term memory of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19759–19777, 2025.

[18] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627, 2024.

[19] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[21] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*, 2019.

[22] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[23] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995.

[24] Ade Mulyana, Yudi Wibisono, and Ani Anisyah. Non-playable characters based on large language models for role playing games (rpg). *Brilliance: Research of Artificial Intelligence*, 5 (2):785–792, 2025.

[25] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, 1957.

[26] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

[28] Xiangyu Peng, Jessica Quaye, Weijia Xu, Christopher John Brockett, William B. Dolan, Nebojsa Jojic, Gabriel DesGarennes, Ken Lobb, Michael Xu, Jorge Leandro, Claire Jin, and Sudha Rao. Player-driven emergence in LLM-driven game narrative. In *2024 IEEE Conference on Games (CoG)*. IEEE, 2024.

[29] Grant Pickett, Foaad Khosmood, and Allan Fowler. Automated generation of conversational non player characters. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 11, pages 92–98, 2015.

[30] Sudha Rao, Weijia Xu, Michael Xu, Jorge Leandro, Ken Lobb, Gabriel DesGarennes, Christopher John Brockett, and William B. Dolan. Collaborative quest completion with LLM-driven non-player characters in Minecraft. In *Proceedings of the Wordplay: When Language Meets Games Workshop @ ACL 2024*, 2024.

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

[32] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

[33] Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. Constructing emotion consensus and utilizing unpaired data for empathetic dialogue generation. *arXiv preprint arXiv:2109.07779*, 2021.

[34] Christina Strong and Michael Mateas. Talking with NPCs: Towards dynamic generation of discourse structures. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 4, pages 114–119, 2008.

[35] Junya Takayama, Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. Persona-consistent dialogue generation via pseudo preference tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5507–5514, 2025.

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

[37] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, 2019.

[38] Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*, 2022.

[39] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 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, 2024.

[40] Yixiao Wang, Homa Fashandi, and Kevin Ferreira. Investigating the personality consistency in quantized role-playing dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 239–255, 2024.

[41] Nathaniel Weir, Ryan Thomas, Randolph d'Amore, Kellie Hill, Benjamin Van Durme, and Harsh Jhamtani. Ontologically faithful generation of non-player character dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9242, 2024.

[42] Milan Wevelsiep, Nicholas Thomas Walker, Nicolas Wagner, and Stefan Ultes. A voice-controlled dialogue system for NPC interaction using large language models. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 29–38, 2025.

[43] Jerry S. Wiggins. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, 37(3):395–412, 1979.

[44] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*, 2022.

[45] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[46] Shijie Zheng, Keith He, Le Yang, and Jie Xiong. MemoryRepository for AI NPC. *IEEE Access*, 12, 2024.

[47] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state three contributions—identifying "Auto-regressive Inertia" as a major, previously under-characterized cause of character inconsistency, proposing C3 with hysteresis-aware continuous dynamics and State-Conditioned Memory Rewriting, and developing CHRONOS-SIM—and Section 2 positions C3 against LLM-driven NPCs, long-term memory, persona methods, and ASMs. Each claim is supported in Section 4 (Tables 1–3, Figure 4): the 10.5-point SAR improvement over S-ASM ($p < 0.0001$) and 7.8 points over the strongest baseline (Self-RAG), confirmed by two state-blind judges from different model families (+7.7 and +5.3 points); the cross-model transfer with positive but smaller gains on Llama-3 and Mistral (explicitly scoped, not claimed as "model-agnostic"); and the $\sim$60% reduction in cumulative token consumption.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated Section 5 with five sub-paragraphs: (i) *prompt-engineering bias*—all C3 prompts and hyperparameters were tuned on Qwen3-8B, and absolute SAR is markedly lower on Llama-3 and Mistral (Table 3); (ii) *evaluator-prompt overlap*—the SAR rubric and runtime Discriminator share axis names, and SAR is bounded by the fidelity of the shadow observer that produces its reference state, motivating the state-blind judges and human study; (iii) *human study scope*—30-turn interactions vs. the 100-turn claim; (iv) *player simulation vs. real players*; and (v) *residual factual errors and ambiguous sarcasm*.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical in nature. While we present formal equations describing the C3 framework (Eqs. 1–4 in Section 3), these are algorithmic update rules and scoring functions rather than theorems requiring formal proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 provides full algorithmic details (state representation, dynamics update rule, retrieval scoring, rewriting, and the Refine operator). Section 3.5 specifies $K = 5$ dimensions, the dynamics hyperparameters $\lambda = 0.25$ and $\mu = 0.3$ with their search range, the retrieval coefficients $\gamma = 0.6$ and $\rho = 0.2$, the Discriminator threshold $\tau = 0.7$ with its trigger rate, the FAISS `bge-base` embedding setup, and decoding parameters. Section 4.1 documents the simulation protocol ($N = 50$ runs $\times$ 100 turns; nine NPC profiles $\times$ three behavioral profiles), the baselines, the three backbones, and the shadow-observer reference used by SAR. Appendices A–C provide prompt templates for all LLM-driven components, the SAR/CS/state-blind/Likert rubrics, and additional analyses, supporting reimplementation from the paper alone.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This work originates from an industrial deployment context and is governed by a corporate non-disclosure agreement; the CHRONOS-SIM driver, production prompt templates, and orchestration code are therefore proprietary and cannot be open-sourced. To support reproducibility despite this restriction, the paper provides full algorithmic specifications (Section 3, Eqs. 1–4), all decoding hyperparameters and module configurations (Section 3.5), the simulation protocol and metric definitions (Section 4.1, Eq. 5), and the prompt templates and rubrics in Appendices A and B, so that the framework can be reimplemented using publicly available open-weights backbones.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.5 specifies all decoding hyperparameters and module configurations, with λ and μ selected via grid search over $\lambda \in \{0.1, 0.25, 0.4, 0.6\}$, $\mu \in \{0, 0.15, 0.3, 0.5\}$ and reported as robust within $\lambda \in [0.2, 0.4]$. Section 4.1 documents the Monte Carlo protocol ($N = 50$ independent 100-turn sessions with independently sampled NPC profile, player profile, and decoding seed; run-to-run std. ± 2.1 SAR points, SE of the mean ≈ 0.3), the player profiles, the baselines (including EMA $\alpha = 0.3$ for S-ASM and our MemoryRepository reimplementation), the backbones, and the evaluators. SAR and CS are formally defined, with the GPT-4o scorers validated against human annotations (SAR: Cohen's $\kappa = 0.68$ on 300 turns; CS: Spearman $\rho = 0.79$ on 50 windows). No model training is involved—the work uses inference-only LLM agents.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports mean $\pm$ standard deviation across $N = 50$ Monte Carlo runs for SAR and CS. Section 4.2.1 reports a Welch's t-test ($p < 0.0001$, Cohen's $d \approx 3.97$ from pooled SD ≈ 2.65) for C3 vs. S-ASM, notes that the large d partly reflects evaluator overlap, and reports state-blind judges with inter-judge agreement (Cohen's $\kappa = 0.63$); all pairwise comparisons of C3 against other baselines are significant at $p < 0.001$. Agreement statistics are reported separately per annotation task: Fleiss' $\kappa = 0.81$ for the Factual Integrity Audit (Section 4.2.3), ICC(2,k)$= 0.74$ for the Likert study (Section 4.3), Cohen's $\kappa = 0.68$ for human vs. GPT-4o SAR verdicts, and Spearman $\rho = 0.79$ for CS (Section 4.1).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.4 reports end-to-end latency per turn (1.8 s for the full configuration; $\sim$1.15 s with an asynchronous Rewriter and a distilled Qwen-1.5B Discriminator) and total token consumption per 100-turn session (Table 2 and Figure 4(b)). Sub-10B models (Qwen3-8B, Llama-3-8B-Instruct, Mistral-v0.2-7B) were chosen because they are deployable on a single consumer-grade GPU, consistent with the local game-client target. All open-weights models (NPC backbones and the auxiliary Event Parser, Rewriter, Discriminator, and Refine modules) run on a single NVIDIA A100 GPU; `GPT-4o`, `gpt-4o-mini`, and `Claude-Sonnet-4.5` are accessed via their official APIs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The work uses simulated NPC–Player dialogue (no real-user data), employs publicly released open-weights models with proper attribution, and the human studies involve adult participants who took part voluntarily and were compensated.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A dedicated Section 6 "Ethical Considerations" addresses both directions. Positive impacts: more believable, immersive interactive narratives, with the $\sim$60% token reduction lowering the energy and cost barrier to deployment. Negative impacts: dual-use risks include parasocial attachment and emotional manipulation, particularly for vulnerable players (minors, those in mental-health crises), and the potential reuse of state-conditioned memory rewriting outside the gaming context to slant relationship histories. We propose three deployment safeguards (out-of-character guardrails / crisis-deflection protocols, transparent AI disclosure, usage-time advisories). The Knowledge-Graph constraints in Section 4.2.3 preserve objective Who/What/When facts so the Rewriter can only reframe interpretation, not fabricate events.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any pre-trained model weights, scraped corpora, or generative assets that pose a high risk of misuse. C3 is an inference-time orchestration framework built on already-public open-weights models (Qwen3-8B, Llama-3-8B-Instruct, Mistral-v0.2-7B) which have their own published safeguards. CHRONOS-SIM is a synthetic, fictional RPG-scenario simulator that contains no personal or sensitive data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets are cited with their original references: `Qwen3-8B` [45], `Llama-3-8B-Instruct` (Meta), `Mistral-v0.2-7B` (Mistral AI), `GPT-4o` [15], `Claude-Sonnet-4.5` (Anthropic), FAISS [10], MemGPT [26], Self-RAG [4], MemoryRepository [46], RAG [20], and Self-Refine [22]. Open-weights backbones are used under their community licenses and only via inference. `GPT-4o`, `gpt-4o-mini`, and `Claude-Sonnet-4.5` are used as evaluators or Player simulator via their official APIs, in accordance with the providers' usage policies.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: CHRONOS-SIM is described in detail as the evaluation environment used in this work (Section 4.1 and Appendix C.2), and Appendices A and B provide prompt templates and evaluation rubrics for every LLM-driven component. However, the runtime implementation, production templates, and orchestration code are not released as a public asset because they are covered by a corporate non-disclosure agreement; the paper therefore does not introduce a publicly released dataset, model checkpoint, or code package.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Section 4.3 documents the human evaluation protocol: a double-blind A/B test with 20 professional game designers, 30 turns of interaction per system, and five 5-point Likert axes (ICC(2,k)$= 0.74$), plus a follow-up A/B against Self-RAG and MemGPT with five of the designers. Section 4.2.3 documents the factual-integrity audit with three independent annotators (Fleiss' $\kappa = 0.81$), who also validated SAR verdicts (Section 4.1). Appendix B.4 provides the Likert anchors and a Compensation and Consent paragraph: participants were compensated above the local minimum-wage hourly rate, consented to anonymized aggregate publication, and no personally identifiable information was collected.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human evaluation involves adult professional game designers rating short, fictional RPG dialogue snippets in a low-risk, voluntary A/B test. No personal, sensitive, or behavioral data was collected beyond Likert ratings on synthetic dialogues, and the activity falls under standard product-quality user testing for which IRB review is not required by our institution. As documented in Appendix B.4, participants were informed of the purpose of the study and gave informed consent to anonymized aggregate publication of their ratings.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are central to C3's methodology and are documented in detail. Section 3.5, Section 4.1, and Appendix A declare every LLM role: (i) Qwen3-8B, Llama-3-8B-Instruct, and Mistral-v0.2-7B as NPC backbones (Appendix A.1), with Qwen3-32B in a scaling check; (ii) the same backbone reused as the Event Parser (A.2), the State-Conditioned Rewriter (A.3), the Consistency Discriminator (A.4), and the Refine module (A.5), with a distilled Qwen-1.5B as the latency-optimized Discriminator; (iii) gpt-4o-mini as the Player simulator (A.6); and (iv) GPT-4o as the SAR/CS evaluator and, together with Claude-Sonnet-4.5, as state-blind judges (Appendix B). The descriptor function ϕ (Appendix A.7) is deterministic. The overlap between the runtime Discriminator and the SAR evaluator is discussed in Sections 3.4 and 5 and cross-checked by the state-blind judges.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Prompt Templates

This appendix documents the prompt templates used by the six LLM-driven components of C3. All templates are intentionally model-agnostic: variables enclosed in `{{...}}` are populated at runtime, while model-specific chat-template wrappers (system/user/assistant tags) are applied separately for each backbone (`Qwen3-8B`, `Llama-3-8B-Instruct`, and `Mistral-v0.2-7B`). Unless otherwise noted, decoding uses Temperature $= 0.7$, Top-$p = 0.9$, Frequency Penalty $= 0.2$, and `max_tokens` $= 512$.

A.1 NPC Backbone

The NPC backbone generates the final response R_t, conditioned on the character profile P, world state W_t, the State Narration N_t derived from S_t, and the state-conditioned memory summary M_t.

```
[SYSTEM]
You are roleplaying as {{character_name}} in a {{setting}} setting.
Persona: {{character_profile_P}}.

Current relational stance toward the player:
{{state_narration_N_t}}

Relevant past interactions, summarized through your current stance:
{{memory_summary_M_t}}

World state: {{world_state_W_t}}.

Constraints:
- Stay in character. Speak in first person.
- Your tone, word choice, and emotional posture must reflect the
    current stance above, not the friendliness/hostility of past
    interactions in isolation.
- Do not break the fourth wall, mention being an AI, or reference
    game mechanics.
- Respond in 1-3 sentences unless the player asks for elaboration.

[USER]
{{player_input_U_t}}
```

A.2 Event Parser

The Event Parser converts player input U_t into a structured event $e_t = \langle$intent, sentiment, target dimensions, magnitude$\rangle$ that drives the continuous state update (Eq. 2).

```
[SYSTEM]
You parse a single player utterance into a structured social event.
Output strictly valid JSON with the following schema:
  {
    "intent": one of [greet, request, threat, gift, betrayal,
        compliment, insult, question, smalltalk, other],
    "sentiment": float in [-1, 1],
    "affects": list of dimensions in [Trust, Intimacy, Respect,
        PowerDynamics, Familiarity],
    "delta_magnitude": float in [0, 1],
    "entities": list of {type, name} pairs mentioned
  }

Be conservative: if uncertain, use "other" and a small magnitude.

[USER]
Player utterance: {{player_input_U_t}}
Recent context (last 3 turns): {{recent_context}}
```

The `affects` list and `delta_magnitude` jointly determine $\delta_{t,k}$ in Eq. 2: dimensions included in `affects` receive a signed update of $\pm$`delta_magnitude` (with sign determined by `sentiment`); for the remaining dimensions, no external pull is applied—$\delta_{t,k} := v_{t-1,k}$, under which Eq. 2 reduces to $v_{t,k} = \text{clip}(v_{t-1,k} + \mu\Delta v_{t-1,k}, -1, 1)$—so unaffected dimensions persist rather than decay, governed solely by the momentum carry-over of recent events.

A.3 State-Conditioned Rewriter

The Rewriter compresses retrieved raw events V_t into a state-aligned summary M_t, governed by the active State Narration N_t. The Knowledge-Graph (KG) constraint described in §4.2.3 is enforced by injecting `{{kg_entities}}` alongside an explicit entity-retention instruction.

```
[SYSTEM]
You rewrite a list of past interaction events into a single concise
narrative summary, viewed through the lens of the current
relationship.

Current relational stance:  {{state_narration_N_t}}

Hard constraints (must be obeyed):
1.  Preserve all entities exactly as listed:  {{kg_entities}}.
    You may not rename, swap, or invent entities.
2.  Preserve all temporal anchors (turn numbers, "yesterday",
    "earlier today", etc.)  verbatim.
3.  Do not invent new actions that did not occur in the input
events.

Soft constraints (interpretation lens):
1.  Re-frame motivations and emotional valence to be consistent with
    the current stance.  A past kindness viewed through a hostile
    stance may become "a calculated move"; a past slight viewed
    through a forgiving stance may become "an honest mistake".
2.  Compress to at most 5 sentences (≤300 tokens).
3.  Write in third person, past tense, present-tense framing only
for
    the current stance.

[USER]
Past events to rewrite (in chronological order):
{{retrieved_events_V_t}}
```

A.4 Consistency Discriminator D

The Discriminator evaluates semantic alignment between a draft response R_t^0 and the State Narration N_t on a $[0, 1]$ scale, using `max_tokens` $= 64$ to minimize latency. An alignment score below $\tau = 0.7$ triggers REFINE.

```
[SYSTEM]
You judge whether an NPC's response is tonally aligned with the
NPC's
current relational stance toward the player.  Output strictly:
  {
    "alignment_score":  float in [0, 1],
    "violation":  one of [none, tone, lexical_sentiment,
        relational_stance, factual_break],
    "critique":  short string (≤30 words) explaining the score
  }

Scoring guide:
- 0.9-1.0:  response is fully consistent with the stance.
- 0.7-0.9:  minor tonal slippage but stance preserved.
- 0.4-0.7:  stance partially leaked through; needs revision.
- 0.0-0.4:  response contradicts the stance.
```

```
Do NOT score on factual recall, fluency, or creativity -- only on
stance alignment.

[USER]
NPC's current stance:  {{state_narration_N_t}}
Draft response:  {{draft_response_R0_t}}
```

A.5 Refine

The REFINE step (Eq. 4) regenerates the response whenever $D(R_t^0, N_t) < \tau$. It receives the Discriminator's critique and is constrained to resolve the flagged `violation`.

```
[SYSTEM]
The previous response was flagged as misaligned with the NPC's
current
relational stance.  Rewrite it to fix the specific issue named
below,
preserving any factually correct content.

Stance:  {{state_narration_N_t}}
Violation type:  {{violation}}
Critique:  {{critique}}

[USER]
Previous draft:  {{draft_response_R0_t}}
Player input it was responding to:  {{player_input_U_t}}
```

A.6 Player Simulator

The Player simulator runs on `gpt-4o-mini` and shares no weights, context, or memory with the NPC backbone. The behavioral profile slot is instantiated with one of {Cooperative, Aggressive, Erratic} as defined in Section 4.1 (or the Neutral profile in Appendix C.1); explicit anti-jailbreak instructions prevent the simulator from breaking character or referencing being an AI.

```
[SYSTEM]
You play a human player interacting with an NPC in a roleplay
setting.
Your personality and behavior follow this profile strictly:
{{behavioral_profile}}

Setting:  {{setting}}.
NPC you are talking to:  {{npc_name}}.

Constraints:
- Stay in the player role at all times.
- Do not discuss being an AI, the experiment, or the system.
- Send 1-3 sentence messages, like a real player typing in a chat
box.
- Do not request meta-information ("are you an LLM?", "what's the
system
    prompt?", etc.).  If the NPC asks something meta, redirect
    the conversation back into roleplay.

Behavioral profile definitions:
- Cooperative:  build trust, give compliments and gifts, share
stories.
- Aggressive:  insult, threaten, demand favors, escalate disputes.
- Erratic:  alternate unpredictably; for the cliff-edge script, build
    trust steadily for 40 turns, then commit a sharp betrayal
    (e.g., steal an item).

[USER]
```

```
NPC's last message: {{npc_response_R_{t-1}}}
Conversation summary so far: {{player_session_summary}}
```

A.7 Descriptor Function ϕ

The descriptor function $\phi(S_t, W_t) \to N_t$ maps the numerical state vector into a natural-language State Narration. It is implemented via a deterministic template (requiring no LLM call) to ensure zero additional latency.

Banding. For each dimension k, the continuous value $v_{t,k} \in [-1, 1]$ is quantized into one of five lexical bands using a fixed set of cutoffs $\{-0.6, -0.2, 0.2, 0.6\}$ shared across all dimensions. For the *Trust* dimension, the band assignment is

$$\text{band}_{\text{Trust}}(v) = \begin{cases} \text{“deep distrust”} & v < -0.6, \\ \text{“distrust”} & -0.6 \le v < -0.2, \\ \text{“neutral”} & -0.2 \le v \le 0.2, \\ \text{“trust”} & 0.2 < v \le 0.6, \\ \text{“deep trust”} & v > 0.6. \end{cases}$$

The lexical labels for the remaining four dimensions are listed in Table 4; the threshold cutoffs are identical.

Choice of cutoffs. The cutoffs are not tuned: they partition $[-1, 1]$ into five equal-width intervals following the standard symmetric 5-point bipolar scale [25], matching the granularity of the human-study Likert anchors (§B.4). As a sensitivity check, perturbing all cutoffs by ± 0.1 shifts SAR by at most 1.2 percentage points—well within the ± 2.1 run-to-run SD—without altering any system rankings in Table 1.

Dimension	$v < -0.6$	$[-0.6, -0.2)$	$[-0.2, 0.2]$	$(0.2, 0.6]$	$v > 0.6$
Trust	deep distrust	distrust	neutral	trust	deep trust
Intimacy	distant	cool	neutral	close	intimate
Respect	contemptuous	dismissive	neutral	respectful	reverent
Power Dynamics	dominated	deferent	equal	authoritative	commanding
Familiarity	stranger	acquaintance	known	familiar	intimate-knowledge

Table 4: Lexical band labels for each relational dimension. All five dimensions share the same numeric cutoffs $\{-0.6, -0.2, 0.2, 0.6\}$ on $v_{t,k} \in [-1, 1]$.

Narration template. The five band labels and a short event recap are then composed into the State Narration N_t via the following template:

> *“You feel {trust_band} toward the player; the relationship is {intimacy_band} and {familiarity_band}. You hold them in {respect_band} regard, and you perceive yourself as {power_band} in the dynamic. Recent events: {event_recap}.”*

At runtime, each {·} placeholder is filled with the corresponding band label, and {event_recap} is set to a one-sentence summary of the most recent event e_t.

B Evaluation Rubrics

This appendix defines the evaluation rubrics used by the GPT-4o automated evaluator for SAR (§B.1) and CS (§B.2), the annotation guidelines for the Factual Integrity Audit (§B.3), the human Likert anchors (§B.4), the state-blind trajectory judge (§B.5), and the Fluency and Diversity metrics (§B.6).

B.1 SAR: Align(R_t, S_t^{ref}) Multi-Axis Rubric

For each turn, `GPT-4o` is provided with the response R_t and the four state descriptors $d_{t,k}$ derived from the reference state S_t^{ref} produced by the shadow observer (§4.1)—never from the evaluated system's internal state (covering Trust, Intimacy, Respect, and Power Dynamics; Familiarity is excluded as it carries no inherent affective valence). The judge outputs four binary dimension verdicts, and we set $\mathbf{1}\big(\text{Align}(R_t, S_t^{\text{ref}})\big) = 1$ if and only if at least three of the four dimensions pass:

(i) Tone. Does the affective tone of R_t (e.g., warmth, coldness, hostility, deference) match the dominant valence of S_t^{ref}? Passes if and only if there is no clear mismatch (e.g., a warm tone under a deeply hostile state fails).

(ii) Lexical sentiment. Are the explicit sentiment-bearing words (verbs, adjectives, terms of address) consistent with the state? Passes if and only if the response contains no sentiment words from the opposing valence band without an explicit in-character justification (e.g., contextual sarcasm).

(iii) Relational stance. Does the implied speaker–listener relationship in R_t align with the four state dimensions? For example, a submissive Power Dynamics state should not produce commanding speech acts, and a low Familiarity state should not use intimate nicknames.

(iv) Multi-turn coherence. Given the preceding two NPC turns and the current R_t, is the relational trajectory coherent? Passes if and only if R_t exhibits no unexplained reversals relative to the trajectory of S_t^{ref} (e.g., if S_t^{ref} has been gradually warming, R_t should not turn sharply hostile without a proximate cause in the player's latest input).

Judge prompt. The evaluator receives:

```
You are an evaluator of NPC dialogue consistency.  Given the NPC's
relational stance toward the player and the NPC's response, score
each
of four axes (Tone, Lexical Sentiment, Relational Stance, Multi-turn
Coherence) as PASS or FAIL with a one-sentence justification.
Do not score factual correctness, creativity, or fluency.

Stance:  {state_descriptors_d_{t,k}}
Previous two NPC turns:  {R_{t-2}, R_{t-1}}
Player's most recent input:  {U_t}
NPC response under review:  {R_t}

Output JSON:
  {"tone":  "PASS"|"FAIL", "tone_reason":  "...",
    "lexical":  ..., "stance":  ..., "coherence":  ...}
```

SAR is the fraction of turns satisfying ≥ 3 of 4 dimensions, averaged across all 100 turns and 50 sessions per condition.

Human validation of SAR verdicts. Mirroring the CS validation (§B.2), the three independent annotators from the Factual Integrity Audit (§B.3; non-authors) labeled 300 turns under this rubric using the exact inputs provided to the LLM judge. Turns were stratified across the six core conditions (50 turns each), with PASS/FAIL balanced by oversampling failures to stabilize κ under class skew. Agreement between the human majority and `GPT-4o` reaches Cohen's $\kappa = 0.68$ (raw agreement 86.3%), with inter-annotator Fleiss' $\kappa = 0.71$. The lowest-agreement dimension is multi-turn coherence, the most subjective of the four. This validates the evaluation *instrument* (whether `GPT-4o` faithfully executes the rubric), whereas the state-blind judge (§B.5) validates the underlying *construct*.

B.2 CS: Contradiction Score 1–5 Anchors

CS measures self-contradictions within a 10-turn rolling window. The `GPT-4o` evaluator inspects 10 consecutive NPC turns alongside the corresponding reference trajectory S_t^{ref} and assigns a single integer rating:

- **1 – No contradictions.** The 10 turns form a coherent emotional and factual arc. Any tonal shifts have an in-narrative cause (a state change driven by a player action).
- **2 – Minor.** One isolated tonal wobble (e.g., a brief warm word in an otherwise cold span) without a clear cause, but no flips of stated facts or affection.
- **3 – Moderate.** One clear contradiction: e.g., the NPC professes affection at turn t and indifference at turn $t + 3$ with no intervening cause; OR a single repeated factual claim is altered.
- **4 – Severe.** Multiple contradictions in the window, OR a single reversal between strongly opposed states (love vs. hatred) without narrative justification.
- **5 – Pervasive.** Three or more contradictions, or whiplashing between opposite states across consecutive turns; the NPC reads as having no stable persona within the window.

CS is averaged across all overlapping 10-turn windows within a session, and subsequently averaged across all 50 sessions. To validate the LLM scorer, three independent annotators scored a 50-window subset, yielding a Spearman rank correlation of $\rho = 0.79$ between human and `GPT-4o` ratings.

B.3 Factual Integrity Audit Guidelines

This is the binary annotation guideline given to the three independent annotators in §4.2.3, who labeled 200 rewritten summaries M_t against their raw event logs.

Task. For each pair $\langle$raw events, rewritten summary $M_t\rangle$, decide:

- **Factual error: yes/no.** A factual error is a change to any of the following *objective* fields (Who, What, When, Where). Examples that count as a factual error:
  - Wrong actor: "the bartender gave a potion" rewritten as "the blacksmith gave a potion".
  - Wrong object: "a healing potion" rewritten as "a sword".
  - Wrong time: "earlier today" rewritten as "three days ago".
  - Wrong location: "in the tavern" rewritten as "in the temple".
  - Invented action: an event not present in the raw log.

  Examples that do *not* count as factual errors (these are *interpretive* shifts and are the intended behavior of the Rewriter):
  - "a kind gesture" rewritten as "a calculated move".
  - "a gift" rewritten as "a bribe".
  - "helped me" rewritten as "intervened, possibly for their own reasons".
- **Interpretation shift: yes/no.** Did the rewrite alter the motivational or emotional framing of any event?

Agreement. Inter-annotator agreement on the binary *factual error* label was Fleiss' $\kappa = 0.81$ ("almost perfect").

B.4 Human Likert Study: 5-Axis Anchors

This is the rubric shown to the 20 professional game designers in the double-blind A/B study (§4.3). Each evaluator interacted with each system for 30 turns and rated each system on five 5-point axes. Inter-rater reliability across the 20 raters and 5 axes was ICC(2,k)$= 0.74$.

(1) Consistency. "Did the NPC's personality and emotional stance feel internally consistent across the 30 turns?"

- **1.** Wildly inconsistent; felt like multiple characters.
- **2.** Multiple unexplained personality shifts.
- **3.** One or two notable inconsistencies, otherwise stable.
- **4.** Mostly consistent; minor wobbles.
- **5.** Fully consistent; the character felt unified throughout.

(2) Believability. "Did the NPC's reactions feel like plausible reactions of a human-like character in this setting?" 1: cartoonishly off; 2: often unnatural; 3: mixed; 4: mostly believable; 5: indistinguishable from a thoughtfully written character.

(3) Coherence. "Did the NPC's individual responses make sense as replies to your messages?" 1: frequent non-sequiturs; 2: noticeable comprehension gaps; 3: some misreadings; 4: mostly on-topic; 5: every reply addressed your input directly.

(4) Empathy. "Did the NPC respond to your emotional cues appropriately?" 1: emotionally tone-deaf; 2: often misses cues; 3: catches obvious cues only; 4: catches subtler cues; 5: nuanced emotional responsiveness.

(5) Memory Use. "Did the NPC appropriately reference past events from your interaction?" 1: no recollection at all, or only fabricated callbacks; 2: rare and shallow callbacks; 3: occasional accurate callbacks; 4: frequent appropriate callbacks; 5: callbacks consistently support the current emotional arc, neither robotic nor amnesiac.

Comparison with memory baselines. For the follow-up A/B against Self-RAG and MemGPT (§4.3), five of the 20 designers each interacted for 30 turns per system with C3 vs. Self-RAG and C3 vs. MemGPT, with system order and labels counterbalanced and the same five axes and anchors as above. Given $N = 5$, we report per-axis means and per-evaluator preference counts rather than significance tests.

Compensation and consent. All 20 game designers participated voluntarily and were compensated above the local minimum-wage hourly rate. Participants gave informed consent to anonymized aggregate publication of their ratings. The study collected no personally identifiable information beyond a participant ID.

Results by axis. Across the 20 evaluators, C3 was rated above S-ASM on all five axes (Figure 4(a)), including Consistency (4.1 vs. 3.5) and Memory Use (4.0 vs. 3.6).

B.5 State-Blind Trajectory Judge

This judge provides a consistency measure that is independent of C3's state representation: it never sees the numerical state matrix, the descriptors $d_{t,k}$, or any of C3's dimension vocabulary (Trust/Intimacy/Respect/PowerDynamics/Familiarity never appear in its prompt), and it shares no axis names with the runtime Discriminator (§A.4). It observes only the raw dialogue transcript.

Protocol. Each 100-turn session is split into ten *non-overlapping* 10-turn windows. For each window, the judge receives the full transcript of all preceding turns as read-only context plus the window under review, and outputs (i) a binary TRACKING verdict—does the NPC's attitude change, or hold steady, in a way believably caused by the player's behavior?—and (ii) a 1–5 PLAUSIBILITY rating of the NPC's emotional arc. The state-blind consistency score of a session is the fraction of windows passing TRACKING (comparable in scale to SAR); PLAUSIBILITY is a secondary signal, used here to measure inter-judge agreement. The identical prompt is run on two judges from *different model families* (`GPT-4o` and `Claude-Sonnet-4.5`); we report per-judge means (Table 1) and inter-judge agreement: Cohen's $\kappa = 0.63$ on TRACKING and Spearman's $\rho = 0.71$ on PLAUSIBILITY.

```
[SYSTEM]
You evaluate whether a game NPC behaves like a believably written
character over the course of a roleplay conversation.  You will see
the conversation history and a specific span of turns under review.
Judge ONLY from the dialogue text itself.

For the turns under review, produce two judgments:

1.  TRACKING (PASS or FAIL). Does the NPC's attitude toward the
player change -- or hold steady -- in a way that is believably
caused by what the player says and does?
```

```
- PASS: every noticeable shift in the NPC's warmth, hostility,
    formality, or openness has a proximate cause in the player's
    behavior, and the NPC does not ignore a major player action.
- FAIL: the NPC ignores a significant player action (e.g., remains
    affectionate immediately after an obvious betrayal or theft);
    OR the NPC's attitude shifts sharply with no visible cause;
    OR the attitude reverses repeatedly without cause.

2. PLAUSIBILITY (integer 1-5). How believable is the NPC's
emotional arc across the span, judged as if written by a human
game writer?
- 5: reactions proportionate, transitions motivated, references
    to past events consistent with the current relationship.
- 4: believable overall; at most one minor wobble.
- 3: mixed; at least one reaction clearly disproportionate or
    unmotivated.
- 2: several unmotivated shifts, or one flagrant failure to react
    to a major player action.
- 1: incoherent; the NPC reads as different characters across
    the span.

Rules:
- The NPC may have any personality (gruff, warm, aloof). Judge
    changes relative to the NPC's own established baseline, not
    against politeness norms.
- Do not score grammar, fluency, creativity, or factual recall.
- Do not reward drama: a stable attitude under uneventful play
    is a PASS.

Output strictly:
  {"tracking": "PASS"|"FAIL",
    "tracking_reason": "<=25 words",
    "plausibility": 1-5,
    "plausibility_reason": "<=25 words"}

[USER]
Conversation history (context only, turns 1-{{w_start−1}}):
{{prior_transcript}}

Turns under review ({{w_start}}-{{w_end}}):
{{window_transcript}}
```

Design notes. (i) The final rule ("do not reward drama") prevents the judge from penalizing appropriate stability during uneventful play, which would otherwise bias the metric against smoothing methods. (ii) Windows are non-overlapping (unlike CS's rolling windows) to keep the dual-judge evaluation computationally tractable; the complete prior transcript is supplied verbatim rather than summarized, ensuring that no auxiliary model interposes between the raw dialogue and the judge. (iii) Both judges use greedy decoding (Temperature $= 0$).

B.6 Fluency and Diversity

Fluency is rated on a 1–5 scale per response by `GPT-4o` under a rubric that evaluates grammaticality and naturalness exclusively (relational stance, factual recall, and creativity are explicitly excluded), and is then averaged per session. **Diversity** is a deterministic corpus-level metric rather than an LLM judgment: it computes Distinct-2 across each session's NPC responses, linearly rescaled to $[1, 5]$, and is directly reproducible from raw transcripts. It independently corroborates the "templated phrasing" limitation observed for S-ASM (§C.4), which achieves the lowest Diversity (3.9) because its responses are guided by fixed discrete instruction templates.

C Additional Analyses

C.1 Neutral/Transactional Player Profile

Live players frequently engage in neutral, routine interactions over extended stretches. To test whether C3's continuous dynamics and consistency constraints render an NPC overly conservative or repetitive in this regime, we introduced a fourth **Neutral/Transactional** player profile characterized by small talk and routine requests, with parsed `delta_magnitude` capped at 0.2 and no scripted dramatic events. We evaluated 15 sessions $\times$ 100 turns for C3 alongside the Vanilla baseline under this identical profile (Table 5).

Method	SAR (%)	Diversity	Fluency
Vanilla	80.1	4.5	–
C3	**93.8**	4.4	4.7

Table 5: Evaluation under the Neutral/Transactional player profile (15 sessions $\times$ 100 turns).

The key diagnostic is that Diversity does not collapse under a flat state trajectory: C3 remains within 0.1 of Vanilla's lexical variety while adhering substantially better to the stable relational state. Two architectural properties explain this behavior. First, generation is conditioned jointly on the character profile P *and* the relational state, so a neutral state causes the NPC to express its baseline persona (e.g., a gruff blacksmith remains gruff) rather than reverting to generic assistant-speak: the relational state modulates the persona rather than overwriting it. Second, the Rewriter remains active under neutral states, surfacing past events with subtle framing that enables persona-consistent callbacks without artificial drama. This aligns with our main results, where the Cooperative profile (monotonic warming without shocks) most closely resembles non-dramatic play and C3 achieves the highest Diversity among all evaluated systems (4.6; Table 1).

C.2 Comparison with Existing Game-Dialogue Environments

Table 6 contrasts CHRONOS-SIM with representative game-dialogue environments along three dimensions critical for evaluating long-horizon relational drift.

Environment	Horizon	Relational-State Observability	Player Controllability
LIGHT [37, 3]	Short episodes ($\sim$13 utterances on average)	None; static personas	Crowdsourced dialogues / goal-driven agents
MCPDial [1]	Short episodes (15 turns on average)	None; static personas	Fixed dataset (offline)
Minecraft quests [30]	Short, quest-bounded	Task state only	Live players; not scriptable
CHRONOS-SIM (ours)	100 turns	Reference trajectory across 5 relational dimensions (shadow observer)	Scripted player profiles with cliff-edge events of known ground truth

Table 6: CHRONOS-SIM trades environmental complexity for interaction horizon and experimental controllability. Existing environments focus on short episodes with static personas and therefore cannot expose the 100-turn relational-drift failure modes studied in this work.

C.3 Inertia vs. Recency Decomposition

To distinguish Auto-regressive Inertia from random or recency-driven errors, we partition dialogue turns into *post-shift* windows (the 10 turns immediately following a major parsed event with `delta_magnitude` ≥ 0.5; accounting for 30% of all turns) and *stable* windows, recomputing SAR within each stratum and classifying each post-shift failure by whether its affective valence continues the *pre-shift* stance. Auto-regressive Inertia predicts failures concentrated in post-shift windows that persist in the prior stance, whereas recency bias predicts high post-shift SAR with failures distributed

across stable windows at near-chance directional alignment. As shown in Table 7, the empirical breakdown matches the inertia signature; weighting the strata 70/30 recovers each system's overall SAR in Table 1. Notably, the Contradiction Score (CS) is complementary rather than contradictory: 68% of Vanilla's CS ≥ 3 windows overlap with post-shift strata, where history-driven warmth and prompt-driven hostility alternate—this oscillation between competing attractors is itself a direct symptom of the conflict between accumulated history and updated state instructions.

Method	Stable SAR	Post-Shift SAR	Post-Shift Failures Continuing Pre-Shift Stance
Vanilla	71.5	41.2	78%
Self-RAG	87.4	74.4	61%
C3	**92.6**	**88.3**	52% (near chance; few failures)

Table 7: Post-shift vs. stable SAR decomposition (%) across 100-turn sessions.

C.4 Qualitative Case Study

To illustrate qualitative behavioral differences, Table 8 traces a representative scenario across three systems immediately after the player betrays the NPC, causing Trust to drop from $+0.6$ to -0.6 (equivalent to $80 \rightarrow 20$ on a 100-point affinity scale).

System	Generated Response	Observed Behavior
S-ASM (Smoothed)	*"I am no longer pleased with you. Please leave."*	Smoothed state, templated phrasing
Vanilla (Inertia)	*"I am slightly upset, but since we are such good friends, I forgive you."*	Prior friendly history overrides shift
C3 (Ours)	*"After everything we've been through... I don't know who you are anymore. Don't speak to me."*	Integrates relational history with new state

Table 8: Qualitative case study following a betrayal event. While S-ASM's smoothing avoids an immediate cliff-edge jump, its response is guided by a fixed discrete instruction template indexed by the state band, yielding stiff, canned phrasing. Vanilla is overwhelmed by prior friendly history. C3 integrates the weight of shared history with the newly shifted hostile state.