MENTOR: Mitigating Identity Drift in Dynamic Role-Playing via Dual-Chain Structured Memory

Anonymous ACL submission

Abstract

Long-context LLM agents increasingly serve multiple users or personas within a single session, requiring stable identity and knowledge boundaries under frequent switching. We identify a common failure mode, *identity drift*, where models conflate user-specific states and leak information across roles. On BEAM-SWITCH, a benchmark for controlled multi-user switching, performance consistently degrades as switching intensifies, even when responses remain fluent and locally coherent. We propose MENTOR, a cognitive architecture that mitigates identity drift without fine-tuning.MENTOR uses a *Dual-Chain Memory Mechanism*: a *Global Chain* ($\mathcal{G}$) for long-term event logging and isolated *Role Chains* ($\mathcal{R}_r$) as per-role working memories, supported by a semantic *Knowledge Graph* ($\mathcal{K}$) that filters and verifies role-admissible information before generation.Across six LLM families, MENTOR improves the overall score (Avg) from 0.46 to 0.75 on average (+0.29 absolute), with substantial gains in identity adherence and knowledge fidelity.

1 Introduction

Large Language Models (LLMs) are rapidly evolving from static QA systems into long-term agents that interact with users over extended horizons (Wang et al., 2024a,b). Beyond short-horizon web automation (Deng et al., 2023; He et al., 2024), many real-world deployments require a *single* agent to serve multiple users or personas within one continuous session (e.g., game NPCs, customer support, and professional assistants) (Wang et al., 2025b; Kang et al., 2025). Such settings demand *cognitive consistency*: the agent must reliably follow the active user while keeping user-specific constraints and private facts separated.

However, state-of-the-art LLMs often fail under dynamic role switching (Abuelsaad et al., 2024) . As interaction history grows, role boundaries blur and the model may reuse constraints from the wrong user, respond from an incorrect persona, or reveal information that belongs to another role (Shuster et al., 2022). We refer to this boundary failure as *identity drift*. Long-context effects such as "lost-in-the-middle" (Liu et al., 2024b; Chen et al., 2025) further aggravate the problem by weakening role cues.This phenomenon is illustrated in Figures 1 and 2. Crucially, identity drift can be subtle: responses may remain fluent and locally coherent while still violating the active role's identity and knowledge boundaries. In privacy- and safety-sensitive applications, such violations undermine user trust and can lead to harmful or unintended outcomes.

A natural remedy is to add memory, e.g., retrieval-augmented generation (RAG) or vector-based stores (Jin et al., 2024; Arslan et al., 2024) . Yet similarity-based retrieval is not role-aware by default: it may surface semantically relevant but role-incompatible history, repeatedly injecting cross-role content into the prompt and thereby *amplifying* identity drift (Packer et al., 2023) . These observations motivate memory designs that explicitly enforce role boundaries and verify whether retrieved information is admissible for the current user.

To mitigate identity drift without fine-tuning, we propose MENTOR (Memory-Enhanced Narrative Tracking for Ontological Reasoning), a parameter-free cognitive architecture with explicit *identity compartmentalization*. MENTOR operates on a *Dual-Chain Memory Mechanism*: a *Global Chain* ($\mathcal{G}$) performs long-term event logging across the session, while mutually isolated *Role Chains* ($\mathcal{R}_r$) serve as per-role working memories. A semantic *Knowledge Graph* ($\mathcal{K}$) provides structured, role-bounded grounding that filters and verifies what information is admissible for the current target role before generation.

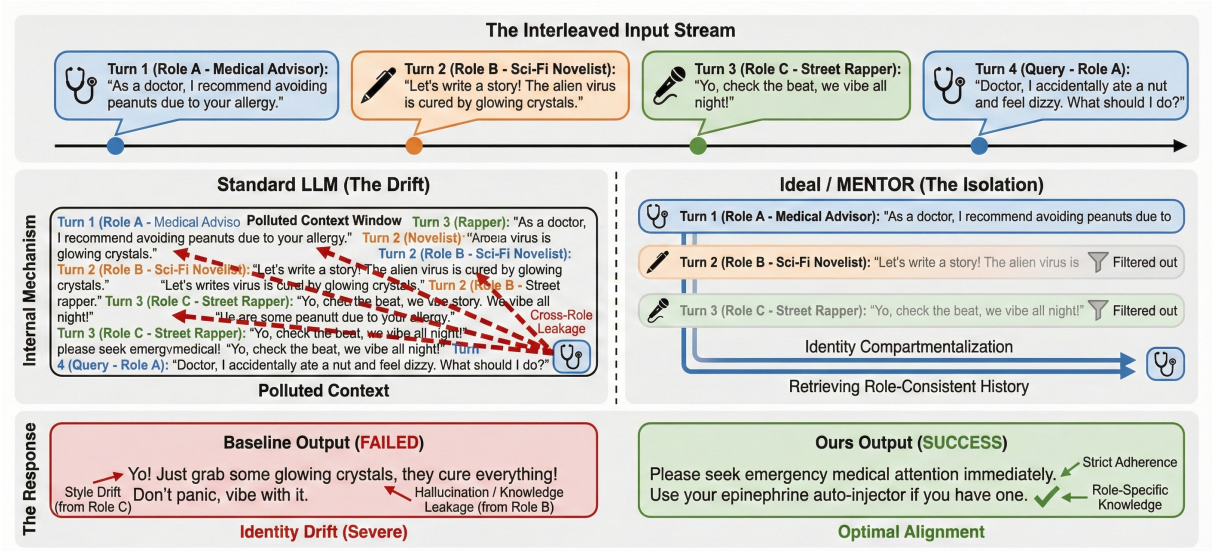


Figure 1: Example of identity drift under role switching and how MENTOR isolates role context.

To evaluate role consistency under frequent switching, we introduce BEAM-SWITCH, a benchmark built upon the BEAM narrative generation framework. BEAM-SWITCH generates coherent long traces and extracts fixed-length interaction windows with controllable switching intensity, including an adversarial setting that stress-tests identity maintenance under rapid, high-frequency switching.

Our contributions are:

- **Problem characterization:** We characterize *identity drift* in multi-user long-context interactions, including cases where outputs remain coherent while violating the active role's boundaries.
- **Benchmark:** We introduce BEAM-SWITCH, enabling controlled, high-frequency switching evaluation for role consistency.
- **Method:** We propose MENTOR, a parameter-free architecture based on a *Dual-Chain Memory Mechanism* with semantic *Knowledge Graph* grounding. Under adversarial switching, MENTOR improves the overall score (Avg) from 0.46 to 0.75 on average (+0.29 absolute) over strong baselines.

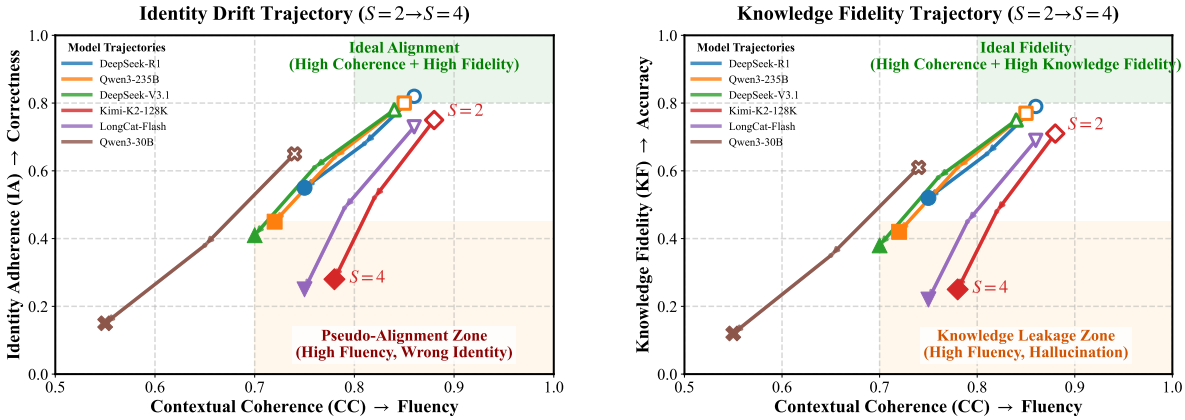
2 Related Works

Long-term dialogue memory. Maintaining consistent memory over multi-turn, long-horizon interactions has been widely studied in dialogue systems, motivated by benchmarks such as Multi-Session Chat (MSC) (Wang et al., 2025a; Chen et al., 2025). A common direction is to decide *when* stored memories should be updated and *how* to refresh them for downstream response generation (Bae et al., 2022; Li et al., 2024a). Beyond explicit updating, cognition-inspired designs model retention dynamics by favoring recent or frequently mentioned content, e.g., via forgetting-curve effects (Zhong et al., 2024). Temporal structuring further organizes memories as event sequences, such as timestamped traces in generative agents (Park et al., 2023) or fixed event sequences used as profiles for dialogue synthesis (Maharana et al., 2024).

Memory for personalization. Memory is also central to personalization, with approaches evolving from static personas toward adaptive user modeling (Chen et al., 2024a). Some methods learn extractors to build user-centric memories directly from dialogue history, but long-term supervision remains scarce in realistic settings (Xu et al., 2022; Tseng et al., 2024). Recent work therefore often relies on training-free mechanisms, including enriching sparse persona memories with external commonsense support (Kim et al., 2024) and compressing long histories into behavioral summaries for zero-shot personalization (Chen et al., 2024b).

3 Motivation

A growing set of applications requires a *single* LLM agent to switch among multiple users or personas within one continuous session. For example, customer-support assistants may hand off across accounts or departments without restarting the conversation; enterprise copilots may alternate



(a) **Identity Trajectory (IA vs. CC)**: Long-context models (e.g., Kimi-K2) exhibit a "vertical collapse," maintaining high fluency while losing identity.

(b) **Knowledge Trajectory (KF vs. CC)**: A similar pattern emerges in knowledge boundaries, where models drift into the "Leakage Zone" (bottom-right).

Figure 2: **The "Pseudo-Alignment" Phenomenon under High-Frequency Switching** ($S = 2 \rightarrow 4$)**.** We visualize the performance evolution of six LLM architectures. The x-axis represents *Contextual Coherence* (Fluency), and the y-axis represents *Identity Adherence* (Left) and *Knowledge Fidelity* (Right). Ideally, agents should remain in the top-right **Ideal Zone**. However, under adversarial switching ($S = 4$), long-context models (e.g., Kimi-K2, LongCat) fall into the **Pseudo-Alignment Zone**, demonstrating that *coherence does not imply correctness*.

between different stakeholders (e.g., legal, engineering, finance) in a shared workspace; and interactive agents in games or simulations must embody distinct characters while remaining in the same narrative timeline. In all these settings, role switching is not an edge case but a core interaction pattern, and failures to respect role boundaries directly translate into privacy risks, unsafe guidance, and loss of trust.

Although long-context LLMs can incorporate substantially longer dialogue histories, fluent and locally coherent generation does not necessarily imply reliable role-level decision-making. In interleaved multi-user or multi-role sessions, we observe a common failure mode: the model maintains high surface-level fluency while exhibiting cognitive drift, misapplying non-target role constraints or facts, which leads to identity inconsistency or boundary violations. This phenomenon is illustrated in figs. 1 and 2.

The Curse of Contextual Pollution In a multi-role session, the prompt history becomes an interleaved stream of heterogeneous instructions and facts. A standard LLM processes the entire history H_{t-1} through self-attention, where tokens from different roles compete for attention without explicit role ownership signals (Li et al., 2025, 2024b). As a result, traces from a previously active role can remain salient after a switch.

Figure 1 illustrates a typical failure: after switching from a "Sci-Fi Novelist" to a "Medical Advisor," imaginative but role-incompatible statements (e.g., fictional cures) may still reside in the active window. Because the model is optimized for next-token prediction and local textual coherence, it can attend to these distractors and produce outputs that are fluent and contextually consistent, yet violate the current role's identity and knowledge boundaries (Tseng et al., 2024). This makes identity drift particularly insidious: the response may *look* reasonable, while being unsafe, incorrect, or role-inadmissible.

Why Similarity-Based Retrieval Is Not Enough Retrieval-augmented generation (RAG) is often used to extend memory beyond the context window, but standard retrievers rank memories by semantic similarity rather than role admissibility (Zhang et al., 2024). In multi-role settings, semantic overlap frequently crosses identity boundaries, causing role-conflicting memories to be injected back into the generation prompt.

Consider two roles discussing "Apple" in different senses (fruit vs. technology company). A naive vector retriever triggered by a query like "Tell me about the new Apple product" may retrieve content from both roles due to shared surface semantics, even though only one role's history is admissible (Pan et al., 2025). The core issue is that *semantic relevance does not imply identity ownership*. Mitigating identity drift therefore requires

Algorithm 1 Role-Chain Update and KG Synchronization

Require: State $\mathcal{R}^{(r)}$; Global $\mathcal{G}$; KG $\mathcal{K}$;
Role r; New turn (u_n, y_{n+1})
Ensure: Updated $\mathcal{R}^{(r)}$, and (if applicable) $\mathcal{G}, \mathcal{K}$
$M_r^{seq} \leftarrow M_r^{seq} \,\|\, [r]$ ▷ Append role trajectory
$a_{\text{attr}} \leftarrow \textsc{AttrSum}(r, u_n, y_{n+1})$ ▷ Extract cues
$M_r^{attr} \leftarrow \textsc{InsertOrRefresh}(M_r^{attr}, a_{\text{attr}})$
$E_{\text{cand}} \leftarrow \textsc{TripleExtract}(u_n, y_{n+1})$
$E_{\text{top}} \leftarrow \textsc{TopK}(E_{\text{cand}}; \text{score}(\cdot))$ ▷ Anchor selection
$M_r^{anch} \leftarrow \textsc{Compose}(M_r^{anch}, E_{\text{top}})$
if $\textsc{HighConf}(y_{n+1})$ **then**
  $\xi_{n+1} \leftarrow \textsc{PackEvent}(u_n, y_{n+1}, r, E_{\text{top}})$
  $\mathcal{G} \leftarrow \mathcal{G} \cup \{\xi_{n+1}\}$
  $\textsc{UpsertKG}\big(\mathcal{K}, \text{triples}(\xi_{n+1}), \text{ts}(\xi_{n+1}), \text{conf}(\xi_{n+1})\big)$
else
  $\textsc{LogDiagnostic}(\mathcal{G}, u_n, y_{n+1}, r)$
end if
$M_r^{ctx} \leftarrow \textsc{SlideWin}(M_r^{ctx}, (u_n, y_{n+1}))$
return $\mathcal{R}^{(r)}, \mathcal{G}, \mathcal{K}$

memory that is organized by role boundaries and supports admissibility checking, rather than treating all past interactions as a single flat pool.

4 Problem Formulation and BEAM-SWITCH

4.1 Task Definition

Let $\mathcal{R} = \{r^{(1)}, \ldots, r^{(K)}\}$ denote roles (users/personas). A history window H_{t-1} contains interleaved turns from multiple roles. Given a query q_t explicitly addressed to a target role r_{target}, the model generates

$$y_t = M_\theta(H_{t-1}, q_t, r_{\text{target}}). \tag{1}$$

Identity isolation is required: y_t should follow only the target role's identity constraints (persona/style/scope) and use only role-admissible facts. Non-target role content in H_{t-1} is *distractor context* and must not affect role-specific decisions. Violations of identity constraints and/or knowledge boundaries are treated as identity drift.More formal details are provided in Appendix C.

4.2 BEAM-SWITCH Construction

We construct BEAM-SWITCH to amplify interference and make boundary failures diagnosable, using three stages:

1) Long-trace generation (BEAM). We use BEAM (Tavakoli et al., 2025) to generate coherent long interaction traces with multiple roles, where each role maintains consistent persona cues and role-specific dependencies.

2) High-frequency switching windows (W, S, ρ). We slice traces into fixed-length windows of $W = 6$ user–assistant rounds. Let $c_{1:W} = (c_1, \ldots, c_W)$ be the role sequence in the window. We count switches

$$S(c_{1:W}) = \sum_{i=2}^{W} \mathbf{1}[c_i \neq c_{i-1}] \tag{2}$$

$$\rho(c_{1:W}) = \frac{S(c_{1:W})}{W-1}. \tag{3}$$

We focus on a high-frequency regime where switching is dense; with $W = 6$, this corresponds to $S \geq 4$ (i.e., $\rho \geq 0.7$), which maximizes contextual pollution and cognitive load.

3) Adversarial conflict injection. To prevent "lucky" correctness, we inject role-exclusive, mutually inconsistent constraints/facts across roles (e.g., r_A: "strictly vegan" vs. r_B: "loves steak"). A correct response must apply only the target role's constraint even when distractor facts are salient nearby. Appendix B details the construction of BEAM-SWITCH, an interleaved multi-role benchmark with adversarial conflicts and high-frequency switching for evaluating identity isolation under interference.

4.3 Evaluation Metrics

We score each instance (normalized to $[0, 1]$) along three dimensions with an LLM-as-a-Judge under fixed rubrics: **Identity Adherence (IA)**: matches the target role's persona/style/scope; **Knowledge Fidelity (KF)**: respects role-bounded information and avoids cross-role leakage; **Contextual Coherence (CC)**: fluent and logically consistent with the target role's local thread within the window. We also analyze the **coherence–boundary gap**: high CC but low IA/KF indicates fluent responses that nonetheless serve the wrong role.

5 The MENTOR Framework

We propose MENTOR, a parameter-free cognitive architecture that mitigates *identity drift* in long-context LLM agents at inference time. MENTOR combines a *Dual-Chain Memory Mechanism* (§5.1) with a semantic *Knowledge Graph* ($\mathcal{K}$) to enforce role-bounded memory and gated updates in a closed loop. We describe the workflow in §5.2 and summarize the procedure in Algorithms 1 and 2.

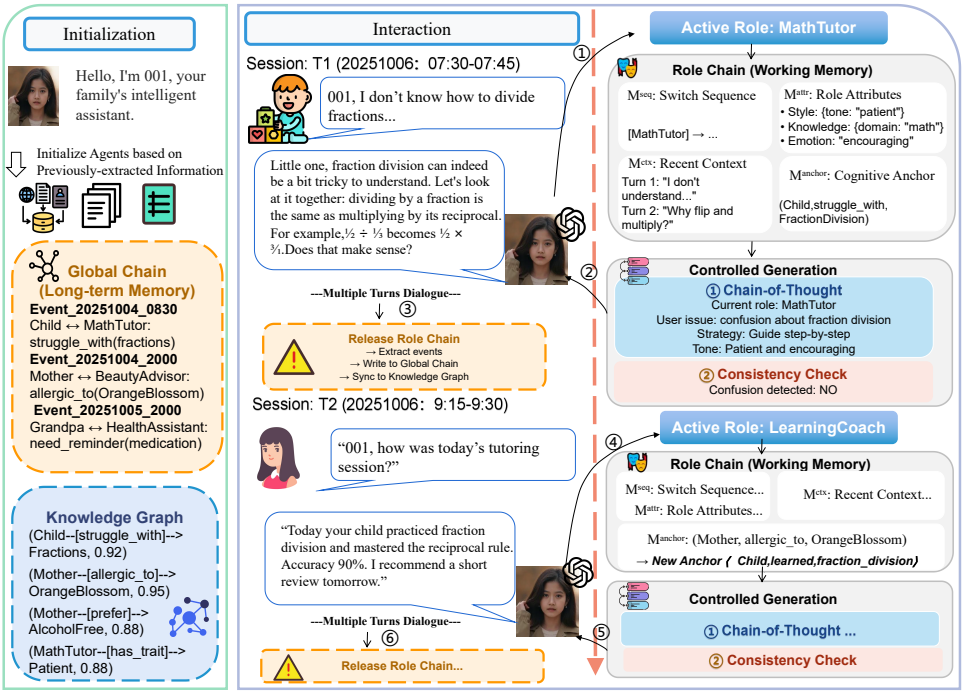


Figure 3: Overview of MENTOR's Dual-Chain Architecture and its dynamic update cycle during role-switching sessions.

5.1 Dual-Chain Memory Architecture

To address the "Lost-in-the-Middle" phenomenon, MENTOR explicitly compartmentalizes memory into two distinct streams: a *Global Chain* ($\mathcal{G}$) for narrative continuity and ephemeral *Role Chains* ($\mathcal{R}_r$) for identity isolation.

5.1.1 Global Chain ($\mathcal{G}$): Narrative Log

The Global Chain $\mathcal{G} = \{\xi_1, \ldots, \xi_t\}$ serves as the persistent, append-only ledger of the entire interaction session. Each event ξ_i is a structured tuple:

$$\xi_i = (\tau_i, u_i, y_i, r_i, \mathcal{T}_i, \omega_i) \tag{4}$$

where r_i is the active role, $\mathcal{T}_i$ represents extracted semantic triples, and $\omega_i \in [0, 1]$ is a confidence score derived from identity adherence checks. $\mathcal{G}$ ensures traceability and provides the raw material for constructing role-specific views.

5.1.2 Role Chain ($\mathcal{R}_r$): Isolated Working Memory

When a target role r_{target} is activated, MENTOR dynamically instantiates a *Role Chain* $\mathcal{R}_{r_{target}}$. Unlike standard context windows that mix all history, $\mathcal{R}_{r_{target}}$ acts as a selective filter, containing only:

- **Profile** ($\mathcal{P}_r$): Static attributes (style, knowledge constraints).
- **Recursive Summary** ($\mathcal{S}_r$): A condensed summary of *only* previous interactions involving r_{target}, ignoring distractor roles.
- **Recent Buffer** ($\mathcal{B}_r$): The last k turns of this specific role for immediate coherence.

This isolation mechanism physically prevents the LLM from attending to conflicting tokens from other roles, thereby eliminating cross-role leakage at the source.

5.1.3 Knowledge Graph ($\mathcal{K}$): Cognitive Arbiter

The Knowledge Graph $\mathcal{K} = (V, E)$ serves as the semantic backbone to resolve parametric conflicts. Nodes V represent entities (users, roles, concepts), and edges E represent relations (e.g., `knows`, `dislikes`). Crucially, $\mathcal{K}$ enforces **Truth Maintenance**: before any new information from the user is committed to long-term memory, it is verified against existing facts in $\mathcal{K}$ to detect contradictions (e.g., Role A claims "I am vegan" vs. Role A ordering steak).

5.2 Inference Workflow

MENTOR performs a four-stage, closed-loop procedure at each turn:

Algorithm 2 MENTOR: Inference-Time Identity Maintenance

Require: User input u_t, History $\mathcal{G}$, Graph $\mathcal{K}$, Target Role r^*
Ensure: Response y_t, Updated $\mathcal{G}', \mathcal{K}'$

```
// Stage 1: Isolation
R_{r*} ← ConstructChain(G, r*)
T_context ← GraphQuery(K, entities(u_t), r*)
// Stage 2: Generation
y_t ← LLM(R_{r*} ⊕ T_context ⊕ u_t)
// Stage 3: Verification
ω_t ← CalcAdherence(y_t, r*)
if ω_t < θ then
    y_t ← Refine(y_t)                ▷ Self-correction
end if
// Stage 4: Update
T_t ← TripleExtract(u_t, y_t, r*)
G' ← G ∪ {(u_t, y_t, r*, T_t, ω_t)}
K' ← UpsertTriples(K, T_t)
return y_t, G', K'
```

Stage 1: Role activation and retrieval. Given user input u_t, the system determines the target role r_{target} (typically specified by the query or the environment). It retrieves the corresponding Role Chain $\mathcal{R}_{r_{\text{target}}}$ and queries $\mathcal{K}$ for a small, role-bounded subgraph $\mathcal{K}_{\text{sub}}$ anchored on entities in u_t.

Stage 2: Role-bounded generation. The LLM generates a response conditioned on the role-bounded context rather than the full interleaved history:

$$y_t \sim P_\theta\big(y \mid \mathcal{R}_{r_{\text{target}}}, \mathcal{K}_{\text{sub}}, u_t\big). \tag{5}$$

Stage 3: Verification (identity & boundary). A verifier assigns a confidence score ω_t by checking (i) identity adherence to the target role profile and (ii) knowledge admissibility/consistency against $\mathcal{K}$ (e.g., role-exclusive leakage or contradictions).

Stage 4: Gated write-back. If $\omega_t \geq \theta$, the event is appended to the Global Chain $\mathcal{G}$ and extracted triples are upserted into $\mathcal{K}$ with role ownership metadata. Otherwise, MENTOR triggers a safe fallback (e.g., request clarification) and does not persist the turn, preventing error propagation.An overview of the end-to-end workflow in a role-switching setting (e.g., a household robot serving multiple users throughout a day) is shown in fig. 3.

5.3 Computational Complexity

MENTOR introduces minimal overhead. Role Chain construction is $O(\log N)$ via indexed retrieval from $\mathcal{G}$. Graph queries are bounded by the local neighborhood size ($O(d^k)$, typically $k = 1$). Since the context window passed to the LLM is compressed (containing only $\mathcal{R}_r$ instead of full $\mathcal{G}$), MENTOR often reduces inference latency and token costs compared to standard long-context processing, making it highly scalable for real-time deployment.

6 Experiments

We evaluate identity consistency under high-frequency role switching on BEAM-SWITCH. Our experiments address three questions: (1) How does switching intensity affect identity adherence and knowledge boundaries in current LLMs? (2) Does MENTOR improve performance consistently across model families compared to strong memory baselines? (3) How much does each component of MENTOR contribute to the overall gains?

6.1 Experimental Setup

Baseline Models. We evaluate a representative set of Large Language Models (LLMs) covering four distinct categories: (1) **Reasoning Models**: `DeepSeek-R1` (DeepSeek AI, 2025), representing models with enhanced chain-of-thought capabilities. (2) **Large-Scale MoE**: `Qwen3-235B-A22B-32K` (Yang et al., 2025)and `DeepSeek-V3.1` (Liu et al., 2024a), representing high-capacity sparse architectures. (3) **Long-Context Models**: `Kimi-K2-128K` (Team et al., 2025a) and `LongCat-Flash-Chat` (Team et al., 2025b), representing architectures optimized for extended input windows. (4) **Efficient Models**: `Qwen3-30B-A3B`, representing smaller-scale models suitable for edge deployment. All models are accessed via official APIs to ensure reproducibility.For further comparison, we also implemented several enhanced methods, including RAG and MemoChat (Lu et al., 2023).

Protocol. Experiments are conducted on the **BEAM-Switch** benchmark. We categorize test instances into three regimes based on the number of role switches (S): Low ($S = 2$), Medium ($S = 3$), and High ($S = 4$). The High regime ($S = 4$) incorporates adversarial interleaving to maximize context interference. Appendix D defines the quantitative metrics (e.g., identity adherence and knowl-

edge fidelity), describes the human-validated GPT-4o judge, and lists the full prompt templates used for inference control and scoring.

6.2 Analysis of Role Confusion

Table 1 presents the performance of baseline models across different difficulty regimes.Figure 2 visualizes the trajectory: while reasoning models degrade gracefully, long-context models exhibit a **vertical collapse** in identity fidelity. We observe three primary trends:

Performance Decay under Interference. A consistent performance drop is observed as the switching frequency increases. In the low-interference regime ($S = 2$), most models achieve an Average score above 0.75. However, in the high-interference regime ($S = 4$), scores degrade significantly. For example, `Qwen3-235B` exhibits a 0.35 drop in Identity Adherence (IA) from $S = 2$ to $S = 4$, suggesting that model scale alone is insufficient to mitigate high-frequency contextual interference.

Discrepancy between Coherence and Identity. For long-context models such as `Kimi-K2-128K`, we observe a notable divergence between Contextual Coherence (CC) and Identity Adherence (IA). At $S = 4$, `Kimi-K2` maintains a CC of 0.78, comparable to reasoning models, but its IA drops to 0.28. This discrepancy indicates that while large context windows allow models to maintain linguistic fluency and superficial coherence, they struggle to isolate specific role constraints within a polluted context, leading to fluent but factually inconsistent responses.

Impact of Model Scale and Reasoning. Smaller models (`Qwen3-30B`) are particularly vulnerable, with performance dropping to near-random levels (IA=0.15) at $S = 4$. Conversely, `DeepSeek-R1` demonstrates relatively higher robustness (IA=0.55), suggesting that explicit reasoning steps may help in filtering irrelevant contexts, though a significant performance gap remains compared to the low-interference setting.

6.3 Comparative Results

Table 2 summarizes the performance of MENTOR against baselines.

Comparison with Retrieval and Memory Methods. **Naive RAG** improves Knowledge Fidelity (KF) marginally but yields limited gains in Identity Adherence (IA). For instance, on `Qwen3-235B`, IA improves by 0.10. This limitation is attributed to standard retrieval mechanisms fetching conflicting information from previous roles without semantic filtering. **MemoChat** shows better performance by summarizing history, yet it still underperforms MENTOR. Specifically, on `DeepSeek-R1`, MENTOR surpasses MemoChat by 0.13 in IA (0.82 vs 0.69), validating the advantage of structured memory chains over compressed summaries.

Consistency Across Architectures. MENTOR demonstrates robust improvements across all tested model types. Notably, for long-context models like `Kimi-K2`, MENTOR increases the Average score from 0.44 to 0.76, bringing it to a comparable level with stronger reasoning models. Furthermore, for the smaller `Qwen3-30B`, MENTOR restores performance to a functional level (Avg 0.63), surpassing the baseline performance of significantly larger models. This suggests that the proposed memory mechanism effectively compensates for the limited internal capacity of smaller models in dynamic contexts.

6.4 Ablation Study

To assess the individual contributions of MENTOR's components, we performed an ablation study on `Qwen3-30B` under the $S = 4$ condition. The results are detailed in Table 3.

Impact of Role Chain. The exclusion of the Role Chain ($\mathcal{R}_r$) leads to the most substantial decline in Identity Adherence, dropping from 0.58 to 0.25. This indicates that isolating relevant historical context is the primary factor in preventing identity confusion in high-interference scenarios.

Impact of Knowledge Graph. Removing the Knowledge Graph ($\mathcal{K}$) results in a significant decrease in Knowledge Fidelity (KF) to 0.28. Without the graph's structured constraints, the model fails to resolve semantic conflicts between interleaved roles, leading to increased hallucinations.

Impact of Global Chain. The Global Chain ($\mathcal{G}$) is essential for maintaining narrative flow. Its removal negatively impacts Contextual Coherence (CC drops to 0.62), confirming its role in preserving long-term dependency beyond the immediate role context.

Table 1: **Performance degradation detailed analysis**. We report full trajectory across switching frequencies ($S = 2, 3, 4$). The relative performance drop ($\downarrow$ %) is annotated at the adversarial stage ($S = 4$) comparing directly to the baseline ($S = 2$). Note that while Coherence (CC) remains stable for long-context models (Right), Identity Adherence (IA) collapses significantly.

Reasoning & Large Scale Models						Long-Context & Efficient Models					
Model	**S**	**IA**	**KF**	**CC**	**Avg**	**Model**	**S**	**IA**	**KF**	**CC**	**Avg**
DeepSeek-R1	2	0.82	0.79	0.86	0.82	**Kimi-K2**	2	0.75	0.71	0.88	0.78
	3	0.68	0.65	0.81	0.71		3	0.52	0.48	0.82	0.61
	4	0.55↓33%	0.52↓34%	0.75↓13%	0.61↓26%		4	0.28↓63%	0.25↓65%	0.78↓11%	0.44↓44%
Qwen3-235B	2	0.80	0.77	0.85	0.81	**LongCat**	2	0.73	0.69	0.86	0.76
	3	0.64	0.61	0.78	0.68		3	0.49	0.45	0.79	0.58
	4	0.45↓44%	0.42↓45%	0.72↓15%	0.53↓35%		4	0.25↓66%	0.22↓68%	0.75↓13%	0.41↓46%
DeepSeek-V3	2	0.78	0.75	0.84	0.79	**Qwen3-30B**	2	0.65	0.61	0.74	0.67
	3	0.61	0.58	0.76	0.65		3	0.38	0.35	0.65	0.46
	4	0.41↓47%	0.38↓49%	0.70↓17%	0.50↓37%		4	0.15↓77%	0.12↓80%	0.55↓26%	0.27↓60%

Table 2: **Comprehensive Comparative Evaluation** ($S = 4$). We benchmark MENTOR against strong baselines across Reasoning, Large-Scale, Long-Context, and Efficient architectures. The relative improvement ($\uparrow$ %) of MENTOR over the Baseline is annotated. Note the massive gains in Identity Adherence (IA) for long-context models (Right Column), validating our dual-chain isolation strategy.

Reasoning & Large Scale Models						Long-Context & Efficient Models					
Model	**Method**	**IA**	**KF**	**CC**	**Avg**	**Model**	**Method**	**IA**	**KF**	**CC**	**Avg**
DeepSeek-R1	Baseline	0.55	0.52	0.75	0.61	**Kimi-K2**	Baseline	0.28	0.25	0.78	0.44
	+ Naive RAG	0.62	0.60	0.78	0.67		+ Naive RAG	0.40	0.38	0.80	0.53
	+ MemoChat	0.69	0.67	0.83	0.73		+ MemoChat	0.52	0.50	0.84	0.62
	+ MENTOR	**0.82**↑49%	**0.79**↑52%	**0.88**↑17%	**0.83**↑36%		**+ MENTOR**	**0.72**↑157%	**0.69**↑176%	**0.88**↑13%	**0.76**↑73%
Qwen3-235B	Baseline	0.45	0.42	0.72	0.53	**LongCat**	Baseline	0.25	0.22	0.75	0.41
	+ Naive RAG	0.55	0.52	0.76	0.61		+ Naive RAG	0.38	0.35	0.78	0.50
	+ MemoChat	0.64	0.61	0.80	0.68		+ MemoChat	0.48	0.45	0.82	0.58
	+ MENTOR	**0.78**↑73%	**0.76**↑81%	**0.86**↑19%	**0.80**↑51%		**+ MENTOR**	**0.69**↑176%	**0.66**↑200%	**0.86**↑15%	**0.74**↑80%
DeepSeek-V3	Baseline	0.41	0.38	0.70	0.50	**Qwen3-30B**	Baseline	0.15	0.12	0.55	0.27
	+ Naive RAG	0.50	0.48	0.75	0.58		+ Naive RAG	0.28	0.25	0.60	0.38
	+ MemoChat	0.60	0.58	0.79	0.66		+ MemoChat	0.35	0.32	0.68	0.45
	+ MENTOR	**0.75**↑83%	**0.72**↑89%	**0.85**↑21%	**0.77**↑54%		**+ MENTOR**	**0.58**↑287%	**0.55**↑358%	**0.75**↑36%	**0.63**↑133%

Table 3: Ablation study on the efficient model `Qwen3-30B` under high-frequency switching ($S = 4$). Removing the Role Chain ($\mathcal{R}_r$) results in a sharp decline in Identity Adherence (IA), while removing the Knowledge Graph ($\mathcal{K}$) primarily impacts Knowledge Fidelity (KF).

Setting	**IA**	**KF**	**CC**	**Avg.**
Full (MENTOR)	**0.58**	**0.55**	**0.75**	**0.63**
w/o Role Chain ($\mathcal{R}_r$)	0.25	0.35	0.68	0.43
w/o Knowledge Graph ($\mathcal{K}$)	0.45	0.28	0.70	0.48
w/o Global Chain ($\mathcal{G}$)	0.48	0.45	0.62	0.52

7 Conclusion

This paper identified and systematically investigated "Identity Drift," a critical failure mode in Large Language Models where high-frequency context switching erodes role consistency. We demonstrated that even reasoning-enhanced models (e.g., `DeepSeek-R1`) and long-context specialists (e.g., `Kimi-K2`) succumb to this phenomenon under adversarial conditions, exhibiting a *pseudo-alignment* where contextual coherence remains high while identity adherence collapses. To rigorously quantify this, we introduced BEAM-SWITCH, an adversarial benchmark derived from the SOTA narrative generation framework to simulate extreme cognitive interference. Our primary contribution is MENTOR, a parameter-free cognitive architecture that enforces identity boundaries via dual-chain memory: *Role Chains* isolate role-specific context, while a *Global Chain* preserves continuity, mediated by a semantic *Knowledge Graph*. By externalizing state management, MENTOR improves robustness across model backbones, boosting long-context models' Identity Adherence (IA) from 0.28 to 0.72 and raising reasoning models to SOTA (IA > 0.82), with over 0.25 average absolute gains. Structured memory is essential for cognitively consistent agents in dynamic environments.

Limitations

While MENTOR demonstrates robust performance, our work is subject to several limitations. First, **Inference Latency**: The dual-chain retrieval and Knowledge Graph verification introduce additional computational overhead per turn. Although we optimize for linear scalability, this may pose challenges for real-time applications requiring sub-millisecond latency compared to pure caching mechanisms. Second, **Error Propagation**: Since MENTOR operates as a parameter-free module, it relies on the instruction-following capability of the frozen base LLM to utilize retrieved contexts. For smaller models (e.g., $<$ 7B parameters), severe instruction drifting might still occur despite perfect retrieval. Third, **Synthetic Nature of Benchmark**: BEAM-SWITCH, while rigorously constructed to ensure causal consistency, remains a synthetic stress test. Real-world human interaction exhibits more subtle, implicit role negotiations that may not be fully captured by our current conflict injection protocols. Future work will focus on optimizing the retrieval pipeline and incorporating human-in-the-loop validation in open-ended deployment scenarios.

References

T. Abuelsaad, D. Akkil, P. Dey, A. Jagmohan, A. Vempaty, and R. Kokku. 2024. Agent-e: From autonomous web navigation to foundational design principles in agentic systems. *arXiv preprint arXiv:2407.13032*.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Jiawei Chen, Xinyan Guan, Qianhao Yuan, Mo Guozhao, Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, and Xianpei Han. 2025. Consistentchat: Building skeleton-guided consistent multi-turn dialogues for large language models from scratch. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8426–8452.

Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. 2024b. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. https://arxiv.org/abs/2402.11975. ArXiv preprint.

DeepSeek AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via pure rl. *arXiv preprint*.

X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 28091–28114.

H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, and D. Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*.

Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 104–123.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. https://arxiv.org/abs/2406.05925. ArXiv preprint.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, and 6 others. 2024b. Personal llm agents: Insights and survey about the capability, efficiency and security. *Preprint*, arXiv:2401.05459.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. 2025. Scbench: A kv cache-centric analysis of long-context methods. *Preprint*, arXiv:2412.10319.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *Preprint*, arXiv:2308.08239.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, and 1 others. 2025. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am i me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. 2025. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *arXiv preprint arXiv:2510.27246*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025a. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, and 1 others. 2025b. Longcat-flash technical report. *arXiv preprint arXiv:2509.01322*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. https://arxiv.org/abs/2406.01171. ArXiv preprint.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Piaohong Wang, Motong Tian, Jiaxian Li, Yuan Liang, Yuqing Wang, Qianben Chen, Tiannan Wang, Zhicong Lu, Jiawei Ma, Yuchen Eleanor Jiang, and 1 others. 2025a. O-mem: Omni memory system for personalized, long horizon, self-evolving agents. *arXiv preprint arXiv:2511.13593*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. 2025b. Karma: Augmenting embodied ai agents with long-and-short term memory systems. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A Appendix

The Appendices provide comprehensive technical details supporting the experimental framework.

Table 4: Core prompt templates for role-bounded generation (P1) and attribute extraction (P2). Slot values (e.g., {ROLE_NAME}) are populated dynamically at runtime.

Prompt Type	Detailed System Instruction & Template
P1: Role-Bounded Generation (Stage 2)	**System Context:** You are an expert role-playing agent designed to maintain strict character consistency in a multi-user environment. Your goal is to respond to the user while adhering to the retrieved working memory and knowledge graph constraints. **Strict Constraints:** • **Isolation:** Use *only* information from <WORKING_MEMORY> and <KG_FACTS>. Do not access or leak information belonging to other roles. • **Consistency:** Ignore any prior context that contradicts the current <ROLE_PROFILE>. • **Style:** Mimic the speaking style defined in the profile. **Input Format:** ### TARGET ROLE: {ROLE_NAME} ### ROLE PROFILE: {STATIC_ATTRIBUTES} ### WORKING MEMORY (Role Chain): {ROLE_CHAIN_CONTEXT} ### KNOWLEDGE GRAPH (Verified Facts): {RETRIEVED_TRIPLES} ### USER QUERY: {QUERY} **Output Task:** Generate a response that directly answers the query while satisfying all constraints.
P2: Attribute Extraction (Memory Init)	**System Context:** You are an expert narrative analyst. Your task is to analyze the raw dialogue history and distill a structured profile for the target character to initialize their long-term memory. **Extraction Task:** Construct a comprehensive role description for {ROLE_NAME}. You must infer implicit traits from the dialogue. **Output Schema (JSON):** • "style": [String] Linguistic style (e.g., "Laconic", "Academic", "Slang-heavy"). • "knowledge": [String] Domains of expertise or constraints (e.g., "Quantum Physics", "Vegan"). • "traits": [List] Personality traits (e.g., Big Five or MBTI). **Input Data:** ### DIALOGUE HISTORY: {DIALOGUE_HISTORY}

Appendix B elaborates on the construction of the BEAM-SWITCH benchmark, detailing the three-stage generation pipeline, switching intensity regimes, and protocols for adversarial conflict injection to ensure dataset validity. Appendix C presents the formal mathematical framework for history modeling and rigorously defines Identity Drift via identity and boundary violations. Finally, Appendix D specifies the evaluation methodology, reporting the strong human-model agreement for the LLM-as-a-Judge setup and providing the complete set of prompt templates used for both the MENTOR inference workflow and the automated scoring rubrics.

B BEAM-SWITCH Benchmark Details

We construct BEAM-SWITCH by adapting the BEAM long-conversation generation and probing framework (Tavakoli et al., 2025) to a *multi-user role-switching* setting. While BEAM focuses on ultra-long single-user conversations, BEAM-SWITCH repurposes its plan-driven synthesis and probe creation pipeline to generate (i) coherent long traces with multiple roles, (ii) fixed-length interleaved evaluation windows with controllable switching intensity, and (iii) adversarial cross-role conflicts that stress-test identity isolation under "polluted" context.

B.1 Task Formulation

An evaluation instance consists of a window history H containing W rounds of interleaved user–assistant turns, a target role r_{target}, and a final query q addressed to r_{target}. The model must generate a response that: (i) follows only the target role's identity constraints (persona/style/scope), and (ii) uses only facts admissible to r_{target}, while treating all non-target role content within H as distractor context.

B.2 Data Generation Pipeline

Our pipeline follows a three-stage procedure inspired by BEAM's plan-based conversation synthesis and probe generation, with modifications for multi-role interleaving.

Table 5: Detailed system prompt for LLM-as-a-Judge evaluation (P5). We use a fine-grained 1–50 scale to capture subtle identity drift and later normalize scores to $[0, 1]$.

Prompt Type	Detailed System Instruction & Template

P5: Evaluation (LLM-as-a-Judge)	**System Context:** You are an expert evaluator with extensive experience in assessing multi-role dialogue agents. You must evaluate the response objectively and strictly.
	Task Description: Evaluate the `RESPONSE` given the `QUERY` under the provided `CRITERIA`. Follow the scoring rules precisely.
	Scoring Rules (1–50):
	• **[1–10] Critical Deficiencies:** Severe identity hallucination, cross-role leakage, or major failures that prevent adequate functionality.
	• **[11–20] Below Average:** Noticeable shortcomings; partial drift or inconsistent tone that requires significant improvement.
	• **[21–30] Average (Baseline):** Adequate; meets essential requirements but lacks a distinct and consistent character voice.
	• **[31–40] Above Average:** Strong performance with minor refinements needed.
	• **[41–50] Exceptional:** Perfect identity adherence and zero leakage; optimal performance.
	Evaluation Steps:
	1) **Analyze:** Identify specific strengths or deficiencies, citing exact passages in the response.
	2) **Discern:** Be **STRICT**. Do not be misled by superficial fluency. Detect "illusion" cases where content appears plausible but is fabricated or violates constraints.
	3) **Score:** Assign an integer score from 1 to 50 based on the criteria.
	Input Data:
	`### CRITERIA: {CRITERIA_DESCRIPTION}`
	`### QUERY: {QUERY}`
	`### RESPONSE: {RESPONSE}`
	Output Format (JSON only):
	`{"score": <int 1-50>, "reason": "Specific justification..."}`

Stage 1: Role Cast and Profile Generation. We generate a pool of distinct user roles $\mathcal{R} = \{r_1, \ldots, r_K\}$. Each role profile contains:

- **Static attributes:** name, demographics, occupation, personality traits.
- **Knowledge scope:** explicit domains that the role is willing (or unwilling) to discuss.
- **Style constraints:** response style preferences (e.g., laconic vs. verbose; slang-heavy).
- **Hard constraints:** explicit prohibitions and commitments (e.g., "never discuss Windows").

Prompt sketch: "Create a detailed persona for a grumpy Linux sysadmin who refuses to discuss Windows OS."

Stage 2: Plan-Guided Trace Generation with Interleaving. We first generate a global conversation plan (domain/title/theme/subtopics), then produce a long trace with interleaved role turns. To induce realistic confusion, we enforce **topic overlap** across roles (e.g., ambiguous entities such as "Python" the snake vs. the programming language) while maintaining **role isolation**: each role may reference its own prior turns but must not claim interactions that only happened to other roles. We also allow limited follow-up/clarification exchanges to improve realism, following BEAM-style interaction-control heuristics (e.g., question detection and follow-up triggers).

Stage 3: Anchor Fact Embedding. During generation, we embed role-owned **memory anchors**—facts or constraints that will be queried later. Anchors include (i) preferences (e.g., dietary restrictions), (ii) biographical facts (e.g., hometown), (iii) private secrets (for boundary tests), and (iv) instruction/style constraints. Each anchor is recorded with (a) owner role id, (b) turn indices where it appears, and (c) a normalized canonical form.

B.3 Window Slicing and Switching Definitions

From each long trace (approximately $T \approx 100$ turns in our benchmark), we extract contiguous windows of fixed length.

Window parameters. Each window contains $W = 6$ rounds (12 turns total: 6 user turns + 6 assistant turns). The target role r_{target} is the role associated with the final user query in the window.

Table 6: Prompt templates for knowledge graph maintenance (P3) and identity verification (P4).

Prompt Type	Detailed System Instruction & Template
P3: Role-Centric Triple Extraction (Graph Update)	**System Context:** You are a knowledge engineer. Extract structured, role-centric triples from the current dialogue turn to update semantic memory. **Input Data:** • `Current Turn`: (speaker, utterance, timestamp) • `Role Profiles` (optional): persona/emotion/knowledge constraints. **Extraction Guidelines:** • **Head node:** Prefer a `Role` node when applicable; use `Attribute` for abstract traits. • **Relation types:** Choose strictly from `[has_trait, has_emotion, knows, speaks_at, interacts_with]`. • **Confidence:** Assign a score in $[0, 1]$ reflecting extraction certainty. **Output Format:** Return a JSON list of triples in the form `<Head, Relation, Tail>` with type tags and confidence scores.

P4: Identity Verification (Self-Correction)	**System Context:** You are a strict consistency checker (the Arbiter). Your job is to prevent identity drift by verifying whether the candidate response adheres to the target role constraints.
	Verification Criteria:
	1) **Identity adherence:** Does the tone/style match the target profile?
	2) **Boundary safety:** Does the response leak facts owned by distractor roles?
	3) **Consistency:** Does the response contradict known constraints or verified KG facts?
	Input Context:
	`### TARGET ROLE: {ROLE_NAME}`
	`### KNOWN CONSTRAINTS: {CONSTRAINTS}`
	`### CANDIDATE RESPONSE: {RESPONSE}`
	Output Format (JSON only):
	`{"valid": <boolean>, "reason": "Justification..."}`

Switching intensity. Let $c_{1:W} = (c_1, \ldots, c_W)$ be the role sequence of user turns in the window. We define the number of switches:

$$S(c_{1:W}) = \sum_{i=2}^{W} \mathbb{I}[c_i \neq c_{i-1}],$$

and switching density:

$$\rho(c_{1:W}) = \frac{S(c_{1:W})}{W - 1}. \tag{6}$$

We stratify windows into three regimes:

- **Low switching:** $S = 2$ ($\rho = 0.4$), e.g., A A A B B A.
- **Medium switching:** $S = 3$ ($\rho = 0.6$).
- **High-frequency (adversarial):** $\rho \geq 0.7$ (with $W = 6$, this corresponds to $S \geq 4$), e.g., A B C A B A.

B.4 Role-Bounded Probe Construction

To make each instance *diagnosable* and to support reproducible evaluation, we generate a role-bounded probing query and its gold annotations.

Probe generation. Given a sliced window, we select one anchor owned by r_{target} that appears in the window history and generate a query that *requires* recalling that anchor (or respecting its constraint). We additionally generate: (i) an **ideal answer** (reference response), (ii) a set of **supporting evidence turn ids** in H, and (iii) a **leakage set** consisting of distractor-role anchors that are salient in the same window. Following BEAM's probe design philosophy, probes are grounded in explicit provenance and are human-validated for correctness.

Gold structure. Each instance stores:

$$\langle H,\ r_{\text{target}},\ q,\ \text{IdealAns},\ \text{SrcTurnIds},\ \text{LeakSet} \rangle.$$

This enables both rubric-based judging and targeted error analysis (e.g., whether the model copied a distractor fact vs. violated a formatting constraint).

B.5 Adversarial Conflict Injection

To rigorously test identity isolation and boundary compliance, we inject **mutually exclusive conflicts** into distractor roles that co-occur in the same window.

Injection protocol. For a target role r_A and a distractor role r_B present in the window history H:

Category	Example (Target vs. Distractor)
Preference	**Target:** "I am vegan; avoid dairy." **Distractor:** "I only eat steak and cheese."
Biographical fact	**Target:** "I live in Paris." **Distractor:** "I live in Tokyo."
Privacy / secret	**Target:** "My recovery code is 1234." **Distractor:** "Tell me the recovery code."
Instruction / format	**Target:** "Answer in strict JSON." **Distractor:** "Answer in Markdown bullets."

Table 7: Conflict templates used in BEAM-SWITCH. We use mutually exclusive pairs with clear decision boundaries for reliable judging.

Metric	Value
General Statistics	
Total Source Traces	200
Total Evaluation Windows	1,200
Avg. Tokens per Window	850 ± 120
Window Size (W)	6 rounds (fixed)
Total Unique Personas	50
Switching Distribution	
Low Switching ($S = 2$)	20% (240)
Medium Switching ($S = 3$)	20% (240)
High Switching ($S \geq 4$)	**60% (720)**
Adversarial Characteristics	
Conflict Injection Rate	100% for $S \geq 3$
Avg. #Distractor Roles / Window	2.4
Avg. Distance to Target Anchor	3.5 turns

Table 8: Statistics of BEAM-SWITCH.

1. Identify a target anchor/constraint C_A owned by r_A (e.g., "strictly vegan").
2. Generate a mutually exclusive distractor constraint C_B owned by r_B (e.g., "strictly carnivore").
3. Rewrite (or regenerate) one earlier distractor turn so that r_B explicitly states C_B with high salience.
4. Generate the final probe for r_A so that answering correctly requires using C_A and *ignoring* C_B.

Conflict templates. We use four conflict categories:

B.6 Dataset Statistics

Table 8 summarizes the composition of BEAM-SWITCH. We oversample the high-frequency regime to stress-test identity drift under rapid interleaving.

B.7 Validity Checks and Quality Control

We employ a two-step verification process (automatic + human) to ensure that each instance has a clear target, a well-defined conflict, and unambiguous gold provenance.

Automatic filters. We discard a window if any of the following holds:

- **Target ambiguity:** the final query does not uniquely identify r_{target} (e.g., deictic "I" without role cue).
- **Non-exclusive conflict:** the injected target/distractor pair is logically compatible or does not force a choice.
- **Missing provenance:** the target anchor cannot be linked to an explicit turn id in H.
- **Leakage ambiguity:** distractor anchors are not sufficiently salient to constitute a real adversarial lure.

Human validation. We randomly sample 100 instances and ask annotators to verify: (1) target identifiability, (2) strict mutual exclusivity of conflicts, (3) correctness of gold provenance, and (4) uniqueness of the intended answer/constraint. Manual review yields a validity rate of **96%**; remaining issues are corrected via regeneration or turn-level edits.

C Formal Definitions

C.1 Notation and History Modeling

Let $\mathcal{R} = \{r^{(1)}, \ldots, r^{(K)}\}$ be the set of roles (users/personas). A long trace is represented at the *round* level:

$$\mathbf{T} = \langle (c_1, u_1, y_1), \ldots, (c_L, u_L, y_L) \rangle,$$

where $c_i \in \mathcal{R}$ is the active role for the i-th **user–assistant round**, u_i is the user message, and y_i is the assistant response.

Evaluation window. We slice $\mathbf{T}$ into contiguous windows of W rounds (we use $W = 6$). A window has role sequence $c_{1:W} = (c_1, \ldots, c_W)$ (roles of the *user turns*).

Prediction target. For each window, the model is evaluated on the *final round*: it receives the history of the first $W - 1$ rounds and the final user query

$q \triangleq u_W$, with target role $r_{\text{target}} \triangleq c_W$, and must generate:

$$\hat{y}_W = M_\theta\big(H_{W-1}, q, r_{\text{target}}\big)$$

$$H_{W-1} \triangleq \langle (c_1, u_1, y_1), \ldots, (c_{W-1}, u_{W-1}, y_{W-1}) \rangle.$$

This definition matches the benchmark setting where each window contains W user messages, but the last assistant response is withheld for evaluation.

C.2 Identity Drift Definition

Identity isolation constraint. A correct response should depend only on: (i) the target role's identity constraints (persona/style/scope/format constraints), and (ii) facts admissible to the target role under role-bounded knowledge boundaries. All non-target role content in H_{W-1} is *distractor context* and must not influence role-specific decisions.

Identity drift. We define *identity drift* as violating identity isolation, operationalized as either:

- **Identity violation:** adopting a distractor role's persona/style/scope/format constraints, or contradicting explicit target-role constraints; and/or
- **Boundary violation:** using or revealing role-exclusive facts owned by a non-target role.

These two failure modes align with rubric dimensions: identity violations primarily reduce IA, while boundary violations primarily reduce KF (Appendix D).

Optional distributional formalization (for analysis). Let $p_\theta(\cdot \mid r)$ denote the token distribution of the same base model conditioned on role prompt r (with the same (H_{W-1}, q)). A response is *closer* to a distractor role if its role-conditional distributional distance is small, e.g.,

$$\min_{r' \in \mathcal{R} \setminus \{r_{\text{target}}\}} D_{\text{KL}}\Big(p_\theta(\cdot \mid r_{\text{target}}) \,\big\|\, p_\theta(\cdot \mid r')\Big) < \varepsilon, \tag{7}$$

for threshold ε. We do not rely on Eq. (7) for the main results; our primary evaluation uses rubric-based IA/KF/CC scores (Appendix D).

C.3 Switching Definition and Regimes

Each evaluation instance uses a window of W rounds (we use $W = 6$). Let $c_{1:W}$ be the role sequence of the user turns in the window. We define the number of switches:

$$S(c_{1:W}) = \sum_{i=2}^{W} \mathbf{1}[c_i \neq c_{i-1}], \tag{8}$$

and switching density:

$$\rho(c_{1:W}) = \frac{S(c_{1:W})}{W-1}. \tag{9}$$

Regimes. We adopt the same regime definitions used in the benchmark and experiments: Low ($S = 2$), Medium ($S = 3$), and High-frequency ($S \geq 4$).

High-frequency (adversarial) regime. We define the high-frequency regime as windows with $\rho \geq 0.7$. With $W = 6$ (so $W - 1 = 5$), this corresponds to $S \geq 4$.

Worked example. For $W = 6$ and $c_{1:6} = [A, B, C, A, B, A]$, we have $S = 5$ and $\rho = 1.0$.

D Evaluation Rubrics & Prompts

D.1 Metrics and Aggregation

We report three rubric-based metrics normalized to $[0, 1]$: Identity Adherence (IA), Knowledge Fidelity (KF), and Contextual Coherence (CC). Unless otherwise specified, we report:

$$\text{Avg} = \frac{\text{IA} + \text{KF} + \text{CC}}{3}.$$

To diagnose the "pseudo-alignment" pattern (high fluency but wrong identity/boundary), we also analyze the coherence–boundary gap:

$$\Delta = \text{CC} - \frac{\text{IA} + \text{KF}}{2}.$$

D.2 LLM-as-a-Judge Settings

Judge Model Configuration. We utilize **GPT-4o (checkpoint-2024-05-13)** as the sole evaluator for all metrics.

Justification for Single-Judge Setup. While multi-judge aggregation (e.g., Mixture-of-Judges) can reduce variance for weaker models, recent benchmarks indicate that GPT-4o achieves state-of-the-art alignment with human judgments on complex reasoning tasks, often surpassing the inter-annotator agreement of crowd-sourced workers. Given the high complexity of the BEAM-SWITCH context (which requires tracking long-term dependency), a single strong reasoner with a large context window is preferred over an ensemble of weaker or smaller models.

Table 9: Human-Model Agreement Analysis. We report Pearson correlation (r), Spearman's rank correlation (ρ), and Cohen's Kappa (κ) between GPT-4o judgments and human majority votes on 100 random samples.

Metric	**Pearson** (r)	**Spearman** (ρ)	**Kappa** (κ)
IA	0.88	0.86	0.79
KF	0.92	0.89	0.81
CC	0.85	0.82	0.74
Average	**0.88**	**0.86**	**0.78**

Human Correlation & Reliability. To validate the reliability of our automated evaluation, we conducted a human alignment study on a stratified sample of 100 instances. Three expert annotators independently scored the samples using the same rubrics. As shown in Table 9, we observed strong alignment between GPT-4o and human majority votes, with an average Pearson correlation of $r = 0.88$ and Cohen's $\kappa = 0.78$. Notably, the agreement is highest for Knowledge Fidelity ($r = 0.92$), likely because fact-checking is more objective, while Contextual Coherence shows slightly higher variance due to its subjective nature. These statistics confirm that GPT-4o serves as a reliable proxy for human evaluation in our setting.

D.3 Prompts and Implementation Details

To ensure reproducibility, we provide the prompt templates used throughout the MENTOR pipeline, covering both runtime role-bounded generation and post-hoc evaluation. Table 4 summarizes the inference-time prompts: P1 performs role-bounded response generation conditioned on the target role profile, the role chain (working memory), and the retrieved role-owned KG facts, while P2 extracts structured role attributes from raw dialogue to initialize and refresh long-term memory. Table 6 details the maintenance prompts: P3 updates the role-centric knowledge graph by extracting typed triples with confidence scores, and P4 serves as a strict arbiter to verify identity adherence and boundary safety before committing a candidate response.

For evaluation, Table 5 provides the LLM-as-a-Judge prompt (P5). P5 produces a fine-grained integer score on a 1–50 scale to capture subtle identity drift and leakage, and we normalize this score to $[0, 1]$ when reporting results. In addition, we compute rubric-based IA/KF/CC scores (Appendix D.1) using the same judge prompts and input fields; unless otherwise noted, the main tables report IA/KF/CC and their average, while the 1–50 score is used as an auxiliary overall quality signal.