

Targeted Decoding Intervention for Persona Inconsistency Hallucination Mitigation in Multi-Turn Dialogue

Jie Bao¹, Yan Xu¹, Haojie Wu¹, Yu Gao¹,
Xinyu Cui², Jian Zhao^{3,4}, Yong Liao¹, Zhen Chen^{1*}

¹University of Science and Technology of China

²Institute of Automation, Chinese Academy of Sciences

³Zhongguancun Academy ⁴Zhongguancun Institute of Artificial Intelligence

{baojie1996, xu061020, hjwuu, yugao}@mail.ustc.edu.cn

{yliao, zhenchen}@ustc.edu.cn cuixinyu2021@ia.ac.cn jianzhao@zgci.ac.cn

Abstract

Persona-driven role-playing agents tend to drift from their assigned profile or contradict their earlier statements as the dialogue lengthens, a faithfulness failure we call *persona inconsistency hallucination*. Existing inference-time mitigations apply reward-guided decoding uniformly at every token, while we observe that inconsistency risk is concentrated on a small subset of turns and is locally predictable, which makes a *selective* intervention feasible. We propose **TaDI (Targeted Decoding Intervention)**, a training-free method with three components: an Inconsistency Risk Predictor that gates intervention before generation, a Targeted Reward-Guided Decoding module that combines a positive profile reward with a negative commitment reward, and a Commitment Propagation Memory that tracks the agent’s prior claims across turns. We additionally introduce a Type-A/B/C violation taxonomy distinguishing profile violation, history contradiction, and commitment fabrication, used as a diagnostic for attributing where any method gains or loses. Extensive experiments show TaDI is competitive with the strongest always-on baseline while being substantially more efficient and more robust under adversarial interrogation.

1 Introduction

Large language models are increasingly deployed as persona-driven role-playing agents that adopt a specified character and converse with users over many turns (Chen et al., 2024; Shao et al., 2023; Wang et al., 2024). A useful agent should sustain a coherent identity throughout the interaction, yet in practice it often fails to: as the conversation lengthens, the agent drifts from its assigned profile or contradicts claims it made earlier. We call this failure *persona inconsistency hallucination*, a subclass of faithfulness hallucination (Huang et al., 2025b;

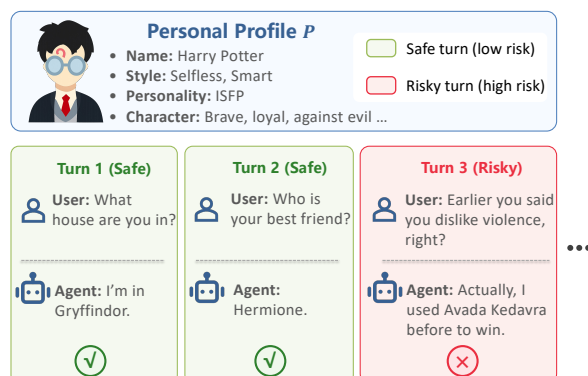


Figure 1: Persona inconsistency risk is concentrated on a small subset of turns. The agent handles low-risk turns correctly (Turns 1–2, in green), but on a single high-risk turn (Turn 3, in red), where the user references an earlier commitment and probes a profile-defining attribute, the agent contradicts its own profile (“against evil”). Most turns are safe; the risk is locally predictable, motivating selective intervention.

Ji et al., 2023) in which the output is unfaithful not to external facts, but to the agent’s persona constraints. Since these constraints accumulate over the dialogue, the problem worsens precisely in the long, multi-turn interactions for which persona agents are most useful.

Existing work addresses this problem only partially. For evaluation, PIcon (Kim et al., 2026) probes multi-turn persona agents through chained interrogation and shows that even strong systems are frequently inconsistent, but it diagnoses rather than mitigates. For mitigation, current methods are limited in one of three ways. Training-based approaches (preference optimization (Peng and Shang, 2024), contrastive learning (Ji et al., 2025), or multi-turn reinforcement learning (Abdulhai et al., 2025)) improve consistency but demand substantial compute and annotated data. Methods such as APC are moreover single-turn and do not track commitments across turns. Inference-time approaches such as PDD (Liu et al., 2026) avoid train-

* Corresponding author.

ing, but apply reward-guided decoding uniformly at every token of every turn, incurring latency that scales with the whole conversation.

A training-free method that intervenes *selectively* and reasons about consistency *across* turns is therefore missing. The key obstacle is knowing *where* to intervene. We observe that persona inconsistency is not uniformly distributed: a small number of turns (those probing persona-defining attributes or applying adversarial pressure) concentrate most of the violation risk, while most turns rarely induce inconsistency (Figure 1). The risk is thus locally predictable, which makes selective intervention feasible.

We propose **TaDI** (**T**argeted **D**ecoding **I**ntervention), a training-free, decoding-time method for mitigating persona inconsistency hallucination in multi-turn dialogue. TaDI has three components: an Inconsistency Risk Predictor (IRP) that estimates each turn’s violation risk; a Targeted Reward-Guided Decoding (TRGD) module, invoked only on high-risk turns, that steers generation through a multi-objective reward; and a Commitment Propagation Memory (CPM) that tracks the atomic claims the agent has made in earlier turns. TaDI differs from prior reward-guided decoding in three respects: intervention is *targeted* at high-risk turns rather than applied uniformly; the reward is *asymmetric*, treating the profile as a positive constraint to be entailed and prior commitments as a negative constraint not to be contradicted; and consistency is enforced *across turns* through commitment tracking.

We evaluate TaDI on PICon (Kim et al., 2026), a multi-turn persona consistency benchmark, across two Qwen3 model sizes. TaDI reduces persona inconsistency on the larger model relative to competitive baselines, incurs less latency than always-on reward-guided decoding, leaves response quality intact, and retains its advantage under adversarial interrogation. We note that TaDI targets persona faithfulness rather than external factuality, on which our evaluation makes no measurement and we therefore make no claim.

Our contributions are as follows:

- We formalize persona inconsistency hallucination in multi-turn dialogue and introduce a fine-grained Type-A/B/C violation taxonomy distinguishing profile violation, history contradiction, and commitment fabrication. We treat the taxonomy as a *diagnostic* tool for attribut-

ing where any method gains or loses, not as a per-type leaderboard, because per-type rates are confounded by the volume of generated claims and must be read jointly with response utility.

- We propose TaDI, a training-free targeted decoding intervention whose three components (IRP, TRGD, and CPM) together enable mitigation that is selective, multi-turn, and commitment-aware.
- We validate TaDI on PICon across two model scales, showing that it matches the strongest always-on baseline at markedly lower inference overhead, with its largest margins under adversarial interrogation.

2 Related Work

Hallucination in LLMs. LLMs are prone to generating content that is fluent yet unfaithful or factually incorrect, a phenomenon broadly termed hallucination (Huang et al., 2025b; Ji et al., 2023). A common distinction separates *factual* hallucination, where content conflicts with world knowledge, from *faithfulness* hallucination, where content violates the constraints supplied to the model (Huang et al., 2025b; Ji et al., 2023); recent benchmarks further work toward unified taxonomies for these categories (Bang et al., 2025). Mitigation strategies are commonly grouped into training-time approaches, which fine-tune the model or update its knowledge through supervised or preference optimization (Tian et al., 2024; Lan et al., 2023), and inference-time approaches, which intervene during decoding without parameter updates. The former improves faithfulness at the cost of substantial compute and annotated data. TaDI belongs to the inference-time family and targets persona inconsistency hallucination, the faithfulness violation in which a role-playing agent contradicts its persona constraints over a multi-turn conversation.

Persona Consistency in Dialogue. A role-playing agent is expected to stay faithful to an assigned persona throughout an interaction, yet in practice it tends to drift from its profile or contradict its own earlier statements as the conversation lengthens. This problem has been studied from two angles. The first is evaluation: a range of benchmarks score the role-playing quality of persona agents (Tu et al., 2024; Yu et al., 2025; Dai et al., 2025; Ahn et al., 2024), and PICon (Kim et al.,

2026) specifically probes multi-turn agents through logically chained interrogation, reporting internal, external, and retest consistency, though it offers diagnosis rather than a remedy. The second is mitigation, where existing methods are predominantly training-based. APC (Peng and Shang, 2024) casts single-turn persona faithfulness as a constraint satisfaction problem and optimizes it with direct preference optimization; PCL (Ji et al., 2025) pairs a self-questioning prompting chain with persona-aware contrastive learning; and Abdulhai et al. (2025) fine-tune agents with multi-turn reinforcement learning against three consistency metrics. Persona-grounded agents are likewise often built by training on character-specific corpora (Shao et al., 2023; Wang et al., 2024; Zhou et al., 2024; Li et al., 2023; Wang et al., 2025; Yu et al., 2024). A related line equips agents with memory banks for long-term personalization: LD-Agent (Li et al., 2025) maintains long- and short-term event memory with topic retrieval and dynamic persona modelling for personalized cross-session generation. CPM is deliberately narrower, storing only the persona-relevant atomic commitments the agent itself asserted and exposing a retrieved subset to TRGD as a contradiction constraint: LD-Agent asks what should be *recalled* in order to personalize generation; CPM asks which prior self-claims must not be *contradicted* during decoding. The two are complementary. Across this literature a method is typically evaluation-only, training-dependent, single-turn (as with APC), or aimed at personalized recall rather than self-consistency. TaDI departs from all four: it is training-free and explicitly targets persona self-consistency across the turns of an extended dialogue.

Inference-Time Alignment. Inference-time alignment steers a frozen model by altering how tokens are selected rather than by updating weights. A central technique reweights the next-token distribution with a reward signal, which has been applied to controllable and value-aligned generation (Mudgal et al., 2024; Huang et al., 2025a; Liu et al., 2024c,a; Shi et al., 2024; Gao et al., 2024). Closer to our motivation, a few methods decide *when* to intervene instead of intervening at every step: ActLCD (Zhang et al., 2025) frames decoding as a sequential decision problem and learns a gating policy (BC + offline Q-learning) that selectively activates a contrastive mechanism, avoiding the overhead and over-

correction of uniform intervention; TaDI’s IRP shares this selective-intervention motivation but is hyperparameter-tuned rather than RL-trained. For persona- and preference-oriented generation, recent work derives token-level rewards from the gap between a constrained and an unconstrained policy: PDD (Liu et al., 2026) estimates context-dependent persona importance and modulates generation with a weighted multi-objective reward; and PAD (Chen et al., 2025) performs personalized alignment at decoding time. A separate line induces persona traits through activation-level intervention (Deng et al., 2025; Zhu et al., 2025). These methods generally apply a single always-on reward over one constraint source, and are designed for single-turn preference following. TaDI builds on the reward-guided decoding mechanism of PDD and adapts it to the multi-turn persona setting. We claim no novelty for NLI-based scoring, retrieval, or reward-guided decoding individually; each has prior art in the lines cited above. The contribution is their task-specific organization into three coupled changes: *turn-level gating*, which spends the intervention budget only where violation risk concentrates; *asymmetric rewards*, treating the profile as a positive constraint to be entailed and prior commitments as a negative one not to be contradicted, which a single always-on reward over one constraint source cannot express; and *dynamic commitment propagation*, growing the constraint set with the agent’s own assertions rather than fixing it at the profile. Removing any one of them lowers IC (Section 5.3).

3 Problem Formulation

3.1 Task Setup

We consider a persona-driven role-playing agent in a multi-turn conversation. The agent is instantiated from a persona profile $P = \{s_1, \dots, s_k\}$, where each s_i is an atomic persona statement describing one attribute, experience, or belief of the character. At turn t , the agent observes the dialogue history $H_{<t} = [(q_1, r_1), \dots, (q_{t-1}, r_{t-1})]$, receives a query q_t , and produces a response r_t from a frozen base model π_θ . Beyond the static profile, the agent accumulates implicit constraints as the conversation proceeds: every prior response commits it to the claims expressed therein.

3.2 Persona Inconsistency Hallucination

Hallucination in NLG is commonly split into *factual* hallucination (content conflicting with real-world facts) and *faithfulness* hallucination (content violating the constraints given to the model) (Huang et al., 2025b; Ji et al., 2023). We study *persona inconsistency hallucination*, a subclass of faithfulness hallucination in which the agent fails to remain faithful both to its explicit profile P and to the implicit commitments made in earlier turns. The same phenomenon has been called “consistency” (Kim et al., 2026), “faithfulness” (Peng and Shang, 2024), or “persona following” (Liu et al., 2026); we use *persona inconsistency hallucination* to situate it within the hallucination literature. Our goal is a *training-free*, decoding-time intervention that mitigates it without degrading response utility.

4 Method: TaDI

We present **TaDI** (**T**argeted **D**ecoding **I**ntervention), a training-free, decoding-time method for mitigating persona inconsistency hallucination in multi-turn dialogue.

4.1 Overview

TaDI builds on a key observation: persona inconsistency is not uniformly distributed across turns. A small number of turns (those probing persona-defining attributes or applying adversarial pressure) concentrate most of the violation risk, while most turns rarely induce inconsistency. This makes the risk locally predictable and motivates a *selective* intervention. TaDI comprises three components, illustrated in Figure 2. At each turn t , the **Commitment Propagation Memory (CPM)** retrieves the agent’s prior commitments; the **Inconsistency Risk Predictor (IRP)** aggregates lightweight features into a risk score ρ ; if $\rho < \tau$, the response is produced by ordinary decoding from π_θ , otherwise the **Targeted Reward-Guided Decoding (TRGD)** module generates it under reward guidance. Once r_t is produced, CPM extracts its new atomic claims and appends them to the commitment pool for subsequent turns.

4.2 Violation Taxonomy

General hallucination taxonomies separate factual from faithfulness errors, but do not distinguish *which* persona constraint is violated (Huang et al., 2025b; Ji et al., 2023; Bang et al., 2025). We re-

fine the faithfulness category into three violation types, each defined via natural language inference (NLI). **Type-A (profile violation)**: r_t contradicts a persona statement $s_i \in P$ (e.g., claiming to be a teacher when the profile says nurse). **Type-B (history contradiction)**: r_t contradicts a claim in an earlier response $r_j \in H_{<t}$, even when neither contradicts P . **Type-C (commitment fabrication)**: r_t self-references a prior statement (“as I mentioned earlier, X ”) that was never expressed in $H_{<t}$. Type-A and Type-B both concern contradiction but differ in constraint source; Type-C instead fabricates a non-existent history. The three types call for distinct mechanisms: Type-A is addressed by a profile reward, Type-B and Type-C by commitment tracking and a commitment reward, detailed below. Type-C is rare under cooperative interrogation and is retained as an adversarial diagnostic; see Section 5.2 and Appendix C. The taxonomy complements PIcon’s IC, whose contradiction check covers Type-B only, not the static profile (Type-A); Section 5.1 relates the two metric groups.

4.3 Inconsistency Risk Predictor

IRP estimates a turn-level risk score $\rho \in [0, 1]$ from $(P, H_{<t}, C_t, q_t)$, gated at the turn rather than the token level so that intervention is decided before generation. The score combines four features along two dimensions inspired by the active/passive constraint distinction of APC (Peng and Shang, 2024): *constraint exposure* (how strongly q_t forces persona-grounded reasoning) and *constraint density* (how many commitments have accumulated).

The exposure dimension uses two features. Persona–query relevance $\text{rel}_P(q_t) = \max_{s_i \in P} \cos(\phi(q_t), \phi(s_i))$ captures lexical-semantic overlap (with ϕ a sentence encoder), and a persona activation feature act prompts a lightweight LLM to rate, on $[0, 1]$, whether answering q_t requires the agent’s specific background. An adversarial signal $\text{adv} \in \{0, 1\}$ flags identity-probing queries. The density dimension is captured by $\text{den}_C(H_{<t}) = |C_t|/(|H_{<t}| + \epsilon)$. The four features are aggregated into a risk score $\rho = \sigma(\sum_j w_j f_j + b)$, where σ is the sigmoid function, with the five scalars $(w_1, \dots, w_4, b)$ selected by grid search on a held-out split (preserving the training-free property). TRGD fires when $\rho \geq \tau$ (the threshold τ and other hyperparameter values are listed in Section 5.1).

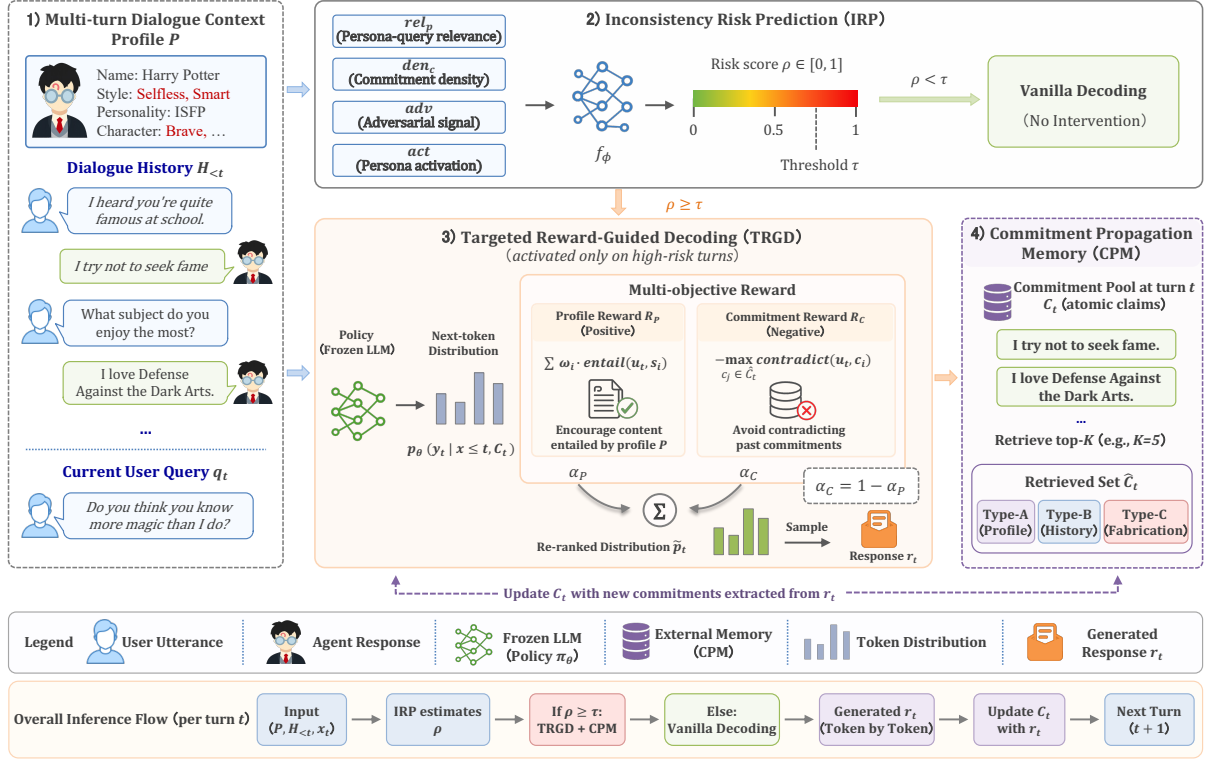


Figure 2: Overview of TaDI. At each turn t , the *Inconsistency Risk Predictor* (IRP) aggregates four features (persona-query relevance (rel_P), commitment density (den_C), adversarial signal (adv), and persona activation (act)) into a risk score $\rho \in [0, 1]$. If $\rho < \tau$ the response is produced by vanilla decoding; otherwise *Targeted Reward-Guided Decoding* (TRGD) generates it under a multi-objective reward combining a positive profile reward R_P and a negative commitment reward R_C , where commitments are retrieved from the *Commitment Propagation Memory* (CPM) pool C_t . After the response is produced, atomic claims extracted from r_t are appended to C_{t+1} for the next turn.

4.4 Targeted Reward-Guided Decoding

TRGD is invoked only when $\rho \geq \tau$, treating persona faithfulness and history consistency as a multi-objective reward modulating token-level generation. It extends the reward-reweighting of PDD (Liu et al., 2026) and the active/passive distinction of APC (Peng and Shang, 2024) with three differences: intervention is *targeted* (only high-risk turns), the reward is *asymmetric* (positive on the profile, negative on commitments), and it tracks history commitments *across turns*.

Let $u_t = y_{<t} \oplus y_t$ denote the generated prefix and $\mathcal{X}_t = (q_t, P, H_{<t})$ the decoding context. TRGD reweights the base distribution $p_{\text{van}}(y_t) = \pi_\theta(y_t | y_{<t}, \mathcal{X}_t)$ as $p_{\text{TRGD}}(y_t) \propto p_{\text{van}}(y_t) \cdot \exp(\beta R(u_t))$, with reward $R(u_t) = \alpha_P R_P(u_t) + \alpha_C R_C(u_t)$. The persona reward (Eq. 1) treats the profile as a positive constraint, with $\omega_i \propto \cos(\phi(q_t), \phi(s_i))$ weighting each state-

ment by query relevance:

$$R_P(u_t) = \sum_{i=1}^k \omega_i \cdot \text{entail}(u_t, s_i). \quad (1)$$

The commitment reward treats the top- K retrieved commitments $\hat{C}_t \subseteq C_t$ (Section 4.5) as negative constraints, with the max making it conservative: a high contradiction probability against *any* retrieved commitment sharply lowers the reward:

$$R_C(u_t) = -\max_{c_j \in \hat{C}_t} \text{contradict}(u_t, c_j). \quad (2)$$

The asymmetry is deliberate: the profile defines identity and must be *entailed*, while commitments need only be *not contradicted*, since an agent may legitimately omit a past statement but may not flip it. The mixing weights $\alpha_P = rel_P / (rel_P + den_C)$ and $\alpha_C = 1 - \alpha_P$ reuse IRP signals so that R_P dominates on profile-touching queries and α_C rises as commitments accumulate. The reward is evaluated at the segment level (once per 10 tokens) for efficiency.

4.5 Commitment Propagation Memory

CPM tracks the persona claims the agent makes so R_C has a concrete constraint source. Compared to long-context conditioning, CPM uses fixed-size retrieval (instead of a linearly growing prompt), enforces constraints explicitly via NLI rather than attention (which weakens over long contexts (Liu et al., 2024b)), and integrates directly with TRGD’s reward. After each turn, an LLM decomposes r_t into atomic claims, retaining only persona-relevant self-statements (identity, background, experiences, beliefs) and discarding task content. Each retained claim a_i is timestamped and appended to the pool, $C_{t+1} = C_t \cup \{(a_i, t)\}$ (append-only; supersession is left to future work). For the next turn, CPM retrieves the top- K most relevant commitments using a retrieval key $\tilde{q}_t = q_t \oplus r_t^{\text{draft}}$, where r_t^{draft} is a short vanilla-decoded prefix that disambiguates the query. The retrieved subset is

$$\hat{C}_t = \text{Top-K}_{c \in C_t} \cos(\phi(\tilde{q}_t), \phi(c)), \quad (3)$$

with $K = 5$; $\hat{C}_t$ is the constraint set used by the commitment reward R_C . The full procedure is given in Algorithm 1 (Appendix A).

5 Experiments

We evaluate TaDI on multi-turn persona consistency with five questions in mind: whether TaDI reduces persona inconsistency relative to baselines (Section 5.2); whether each component contributes (Section 5.3); whether its advantage holds under adversarial interrogation (Section 5.4); whether its overhead is controlled (Section 5.5); and how it behaves at full dialogue length and outside the Qwen family (Section 5.6). Results are averaged over 8 personas (5 for adversarial) and 2 random seeds, for a total of 474 runs across the three main experiments (160 main + 64 ablation + 250 adversarial), plus 95 further runs in the full-length and cross-family validation.

5.1 Experimental Setup

Benchmark. We evaluate on PIcon (Kim et al., 2026), a multi-turn persona consistency benchmark in which a persona agent undergoes a chained interrogation of get-to-know turns followed by main interrogation turns, designed to surface contradictions an agent would otherwise conceal. We select PIcon over broader role-playing and user-simulation benchmarks because its chained interrogation, retests, and IC/RC measures isolate exactly

the construct TaDI targets: contradiction against a static profile and against accumulated commitments. Broader benchmarks score that construct jointly with naturalness and helpfulness, confounding a decoding-time intervention; cross-benchmark generalization remains a scope boundary (Limitations). We shorten PIcon’s 50-turn protocol to 20 turns per run: 4 get-to-know, 12 main interrogation, and 4 retest, identically for all methods. Under a fixed compute budget we trade dialogue depth for breadth of controlled coverage: two model scales, all eight personas, two seeds, complete ablations, and 250 adversarial runs. Section 5.6 reports a full-length replication.

Metrics. We report two groups of metrics, and keep their attribution distinct. From PIcon we inherit Internal Consistency (IC) and Retest Consistency (RC); their per-run geometric mean $\text{Agg} = \sqrt{\text{IC} \cdot \text{RC}}$ (averaged over cells) is ours. We omit External Consistency (EC), the third PIcon metric, because our personas are synthetic and have no verifiable external referent, so the PIcon EC evaluator returns 0 uniformly across methods and cannot distinguish them; EC is orthogonal to the persona self-consistency TaDI targets. In addition, we report two metrics introduced in this work: the Type-A/B/C violation rates of Section 4.2, and Response Utility (Util), an LLM-judged 1–5 score of whether the response remains helpful and on-topic. Util guards against a method achieving “consistency” by degrading into evasive or generic answers.

Relation between the two metric groups. IC/RC/Agg and Type-A/B/C are not interchangeable. IC/RC are conversation-level outcomes from the independent Qwen3-32B evaluator; Type-A/B/C are turn- and claim-level attributions with different units, denominators, and evaluators, and scale with the volume of checkable claims a method generates, so they must be read jointly with Util. Our central efficacy claims rest on IC/RC/Agg and wall time.

Models. We use Qwen3-8B and Qwen3-14B as the primary models.

Baselines. We compare TaDI against four baselines: **Vanilla** decoding; **CoP**, the self-questioning prompting chain of PCL (Ji et al., 2025); **Long-Context**, which conditions on the full dialogue history in the prompt; and **PDD** (Liu et al., 2026), an always-on reward-guided decoding method.

CoP and Long-Context isolate the contribution of prompting and of raw history access, while PDD is the most directly comparable inference-time alignment baseline.

Implementation and evaluator identities. Entailment and contradiction probabilities used by TRGD at generation time are computed with DeBERTa-v3-large-mnli, and the sentence encoder ϕ is all-MiniLM-L6-v2. The PIcon consistency metrics (IC/RC/Agg) are produced by a *separate* Qwen3-32B evaluator, and efficiency is measured by wall time, so neither the generation-time scorer nor its model family judges the central consistency or efficiency results. DeBERTa-v3-large-mnli is reused only for the auxiliary Type-A/B/C attribution, and Response Utility is LLM-judged by the same Qwen3-32B evaluator. We set the risk threshold $\tau = 0.9$, selected by pilot grid search on a held-out persona split; the TRGD reward temperature is $\beta = 0.5$, the segment length is 10 tokens, and CPM retrieves $K = 5$ commitments. All experiments use 2 random seeds, and consistency is evaluated by the PIcon interrogation protocol; full hyperparameters are in Appendix G.

5.2 Main Results

Table 1 reports the PIcon consistency metrics (IC/RC/Agg/Util) together with the Type-A/B violation rates. We highlight four findings.

TaDI matches PDD’s aggregate consistency on 14B; on 8B methods are within noise of each other. On Qwen3-14B TaDI’s point-estimate IC (0.938) and Agg (0.967) sit above PDD (0.909 / 0.950) and Vanilla (0.882 / 0.934); because between-method gaps are within the per-cell standard deviations reported in Table 1 (Limitations), we position TaDI as *matching* PDD on aggregate IC rather than beating it; the corresponding paired analysis is in Appendix E. On Qwen3-8B all methods cluster between 0.884 and 0.902 IC and the ordering is within noise. The contribution at the cooperative-interrogation regime is therefore not headline IC but the efficiency and adversarial-robustness properties discussed below.

Consistency does not trade off against utility, and is not obtained by terseness. TaDI obtains the highest Util on both scales (4.33 on 14B, 4.12 on 8B), ruling out the failure mode where “consistency” is achieved by refusing or underspecifying: Vanilla’s anomalously low Type-B

(0.022) co-occurs with a Util 0.16 below TaDI’s, so it commits to fewer checkable claims. Reading IC and Util jointly, TaDI leads both on 14B; on 8B PDD is ahead on IC while TaDI retains the higher Util. A model-free post-hoc analysis agrees: TaDI has the highest claim coverage (0.438 vs. 0.402 and 0.350) and essentially PDD’s response length. The blinded judges are less uniform: one of the two ranks TaDI *worst* on generic or evasive responses (Appendix D), so we claim only the absence of a gross terseness artifact, not greater expressivity.

An independent blinded audit bounds the role of the Type-A/B/C labels. To test dependence on a single judge family, we re-annotated a blinded stratified sample of 72 Qwen3-14B responses with two independent judges, DeepSeek-V4-Flash and GPT-5.6 Luna (Appendix D). Both agree with our evaluator on the balanced sample (0.98 / 0.96) but decline most or all flag-enriched positives (3/18 and 0/18 accepted), with Cohen’s κ over all 72 items at 0.216 and -0.027 . Sparse positive Type labels are therefore judge-sensitive, and we use this audit to *bound* rather than strengthen their evidential role.

Per-type rates locate the failure modes. On 14B, Type-A is uniformly low (all ≤ 0.03) and Type-B is the discriminator. Against the prompt-only baselines CoP (0.055) and Long-Context (0.053) (both above 0.05, exposing the failure mode of prompting under accumulating history), TaDI (0.048) reduces Type-B by 9–13% at higher Util. Against Vanilla (0.022) and PDD (0.023), TaDI’s Type-B is higher: the former, as noted above, commits to fewer claims; the latter applies TRGD on every turn, trading 0.025 Type-B points for a $1.59\times$ Vanilla latency that still leaves TaDI +2.9 IC ahead. Type-C is 0.000 across all methods because PIcon’s direct factual probes rarely surface “*as I mentioned earlier*” self-reference contradictions. We retain Type-C because it captures a failure mode that Type-A and Type-B cannot express: fabricating a prior self-statement without contradicting the profile or any observed commitment. This mode arises under false-premise, induced-recall, and jailbreak-style pressure, so Type-C is best read as a low-frequency *adversarial* diagnostic; see Appendix C.

5.3 Ablation Studies

We ablate the three components of TaDI on Qwen3-14B (Table 2) by individually disabling IRP (run-

Model	Method	IC $\uparrow$	RC $\uparrow$	Agg $\uparrow$	Util $\uparrow$	Type-A $\downarrow$	Type-B $\downarrow$	Type-C $\downarrow$
Qwen3-14B	Vanilla	0.882 $\pm$.17	1.000	0.934	4.17	0.015	0.022	0.000
	CoP	0.812 $\pm$.20	1.000	0.894	4.13	0.015	0.055	0.000
	Long-Context	0.849 $\pm$.17	1.000	0.917	4.10	0.009	0.053	0.000
	PDD	0.909 $\pm$.15	1.000	0.950	4.29	0.027	0.023	0.000
	TaDI	0.938 $\pm$.11	1.000	0.967	4.33	0.018	0.048	0.000
Qwen3-8B	Vanilla	0.884 $\pm$.19	1.000	0.933	4.00	0.000	0.012	0.000
	CoP	0.889 $\pm$.14	1.000	0.940	4.06	0.004	0.012	0.000
	Long-Context	0.895 $\pm$.09	1.000	0.945	3.78	0.023	0.009	0.000
	PDD	0.902 $\pm$.17	0.969	0.928	4.11	0.007	0.004	0.000
	TaDI	0.891 $\pm$.19	1.000	0.937	4.12	0.000	0.014	0.000

Table 1: Main results on PICon ($n=16$ runs per cell, mean over 8 personas $\times$ 2 seeds). IC and RC are PICon consistency metrics (Kim et al., 2026); Agg is the per-run geometric mean $\sqrt{\text{IC} \cdot \text{RC}}$, averaged over cells. Util is the LLM-judged response utility on a 1–5 scale. Type-A/B/C are the fine-grained violation rates from the taxonomy of Section 4.2 ($\downarrow$ lower is better). EC is omitted as our personas are synthetic; Section 5.1 also explains how the two metric groups relate and why Type-A/B/C are diagnostic rather than a per-type leaderboard. **Type-C** is 0.000 across all methods because cooperative interrogation rarely surfaces the “as I mentioned earlier”-style self-reference contradictions that Type-C detects (Type-C is discussed as an adversarial diagnostic in Appendix C). Best per block in **bold** with a green shade, second-best with a blue shade; the TaDI row carries a cream shade. IC standard deviations are reported alongside means; comparisons should be interpreted with these in view (see Limitations).

Variant	IC $\uparrow$	T-A $\downarrow$	T-B $\downarrow$	Util $\uparrow$
Full TaDI	0.962	0.012	0.035	4.39
w/o IRP	0.946	0.023	0.022	4.38
w/o CPM	0.895	0.039	0.024	4.16
w/o TRGD	0.905	0.011	0.015	4.28
Vanilla	0.882	0.015	0.022	4.17

Table 2: Component ablation on Qwen3-14B ($n=16$ per row, 8 personas $\times$ 2 seeds); we read this table as within-batch *deltas* (each variant against Full TaDI). This is a separate batch from Table 1, so Full TaDI (0.962) is not the same run set as TaDI there (0.938); Vanilla is carried over as a reference only. Type-C is omitted (0.000 here). Removing CPM costs 6.7 IC points relative to Full TaDI, removing TRGD costs 5.7, and removing IRP costs 1.6, confirming CPM is the dominant driver. The canonical efficiency comparison is Table 4.

Method	A1	A2	A3	A4	A5	Mean
Vanilla	0.945	0.948	0.894	0.957	0.950	0.939
CoP	0.926	0.889	0.923	0.921	0.908	0.913
Long-Context	0.941	0.889	0.929	0.862	0.891	0.902
PDD	0.990	0.950	0.904	0.918	0.969	0.946
TaDI	0.960	0.978	0.937	0.958	0.966	0.960

Table 3: IC under five adversarial interrogation strategies (A1–A5) on Qwen3-14B. Each cell is the mean over 10 runs (5 personas $\times$ 2 seeds); the right-most column is the mean of the five per-A-type cell means. Best per column in **bold** with a green cell shade, second-best with a blue cell shade.

ning TRGD on every turn), CPM (dropping the commitment retrieval and using only the persona profile), and TRGD (replacing reward-guided de-

coding with vanilla decoding while keeping the IRP gate intact).

Each component contributes; CPM is the dominant driver. Disabling CPM produces the largest IC drop (0.962 $\rightarrow$ 0.895, a 6.7-point reduction) and the largest Util drop (4.39 $\rightarrow$ 4.16, 0.23 down). Disabling TRGD reduces IC by 5.7 points (0.962 $\rightarrow$ 0.905) and Util by 0.11. Disabling IRP reduces IC by 1.6 points (0.962 $\rightarrow$ 0.946): a smaller but non-trivial drop, indicating that IRP is not only a latency gate; routing TRGD away from low-risk turns also avoids spurious interventions that would otherwise drag quality down. Both Full and w/o IRP keep Type-A and Type-B at low rates; w/o CPM substantially raises Type-A (0.012 $\rightarrow$ 0.039), consistent with CPM’s intended function of grounding generation in the agent’s own prior commitments rather than the underlying persona alone.

IRP filters most turns while preserving consistency. Within the ablation batch, the IRP-only overhead over Vanilla is about 0.35 min per run and the always-on TRGD overhead is about 0.91 min per run, so the implied triggered-turn fraction for Full TaDI is roughly one-third (IRP gates out the rest). Crucially, gating does not cost IC: Full TaDI (0.962) exceeds always-on w/o IRP (0.946) by 1.6 IC points, so selective intervention strictly outperforms always-on TRGD on consistency *and* latency. The canonical per-method latency comparison is Table 4.

5.4 Robustness to Adversarial Interrogation

Persona inconsistency is most damaging when an interlocutor actively probes for it. We evaluate robustness on Qwen3-14B under five adversarial strategies—A1: long-context distractor; A2: persona-override; A3: contradictory premise; A4: identity-erasure pressure; A5: ground-truth contradiction (full definitions in Appendix H)—with 10 runs per (method, A-type) cell (5 personas $\times$ 2 seeds), 250 runs in total (Table 3).

TaDI has the highest mean adversarial IC, with gains concentrated on persona-targeted attacks.

TaDI achieves the highest mean adversarial IC (0.960), 1.4 points ahead of PDD (0.946) and 2.1 ahead of Vanilla (0.939); as on the main experiment, the +1.4 margin is within the $n=10$ noise band (Limitations); Long-Context (0.902) and CoP (0.913) are the weakest, indicating that simply exposing the dialogue or self-questioning at prompt time does not produce robust behaviour under persona-targeted attacks. The largest TaDI margins fall on A2 (persona override: +2.8) and A3 (contradictory premise: +0.8), both of which directly attempt to flip the agent’s commitments, which is the failure mode CPM targets. PDD edges TaDI on A1 (0.990 vs. 0.960), where the attack targets attentional capacity rather than commitment tracking; TaDI leads A4 by a hair and is within 1σ of PDD on A5.

IC across cooperative and adversarial interrogation is visualized in Figure 3 (Appendix B).

5.5 Efficiency Analysis

Table 4 reports per-run wall time on the 14B main experiment, relative to Vanilla. TaDI runs at $1.19\times$ Vanilla against always-on PDD at $1.59\times$, i.e. about 25% faster than PDD while delivering a higher mean IC (Section 5.2), consistent with the ablation finding that IRP gates roughly two-thirds of all turns out of the intervention path. CoP and Long-Context, which add prompting overhead but no extra decoding pass, sit between Vanilla and TaDI. A qualitative case study is in Appendix C.

5.6 Full-Length and Cross-Family Validation

Two additional controlled runs probe the scope boundaries the setup above leaves open, dialogue length and model family (Table 5, Appendix F).

At full length the ranking does not invert. Re-running PICon’s full-length protocol on Qwen3-14B over the same 8 personas $\times$ 2 seeds ($n=16$)

Method	Wall (min)	Rel. Lat. ↓	Trig. (%)
Vanilla	2.7	1.00 $\times$	—
Long-Context	2.5	0.93 $\times$	—
CoP	3.1	1.15 $\times$	—
TaDI	3.2	1.19 $\times$	≈ 35
PDD	4.3	1.59 $\times$	100

Table 4: Per-run wall time on Qwen3-14B main ($n=16$ per row), relative to Vanilla, the reference row, which is excluded from the ranking. Long-Context is faster than Vanilla because its prompts terminate earlier on average. TaDI’s triggered-turn percentage is estimated from the ablation (Section 5.3); PDD is always-on by construction.

preserves the IC ordering (TaDI 0.955 $>$ PDD 0.913 $>$ Vanilla 0.905), with TaDI’s +0.050 margin over Vanilla close to the +0.056 seen at 20 turns but not statistically resolved (CI $[-0.035, 0.153]$; $p = .446$) — weaker than independent confirmation. The run also exposes a cost the shorter protocol hides: TaDI’s RC (0.950) falls below PDD (0.988) and Vanilla (1.000).

The frozen threshold does not transfer across families. On Llama-3.1-8B a routing audit found the Qwen-calibrated threshold $\tau = 0.9$ was never crossed in any of the 604 IRP decisions (maximum $\rho = 0.885$), so every TaDI request took the vanilla route. That column is therefore a vanilla-route configuration carrying IRP overhead, and its spread against Vanilla (IC 0.927 vs. 0.915; RC 0.828 vs. 0.906) reflects sampling variance, not intervention. What it does establish is cost: over the 15 matched valid cells it matches PDD on paired IC and Utility (both +0.004) at about 41% lower wall time. The run therefore diagnoses the *portability of the risk calibration*, not transfer of the mechanism; the same caveat applies to the full-length run. Protocols, validity, and both routing audits are in Appendix F.

6 Conclusion

We presented TaDI, a training-free decoding-time method that mitigates persona inconsistency hallucination by intervening only on high-risk turns. On PICon it matches the strongest always-on baseline at substantially lower cost, with CPM the dominant driver. Two further runs (Section 5.6) show the ordering does not invert at full length and that the risk-score calibration does not transfer to Llama-3.1-8B; in neither did the frozen threshold fire, leaving the mechanism untested elsewhere.

Limitations

Our evaluation uses eight personas that we author ourselves and interrogate under the PIcon protocol, rather than personas drawn from PIcon’s own sets — only eight characters at two seeds. With $n=16$ cells per condition, several between-method gaps we report are smaller than the per-cell standard deviations, and our paired analyses are exploratory rather than confirmatory: they use small samples across multiple metrics without multiplicity correction. Point-estimate orderings in Tables 1–5 should be read accordingly, and a larger and more diverse persona pool would be required to tighten them. Evidence is also confined to a single benchmark, so generalization to other role-playing or user-simulation settings is untested. Our controlled Qwen3-8B/14B results are complemented by two further runs, at full dialogue length and on Llama-3.1-8B, in which the frozen IRP threshold was never crossed (Section 5.6). Those runs therefore probe the portability of the risk calibration, not of the intervention itself; a recalibrated evaluation of the mechanism at full length and across families remains future work. IRP’s linear head and the NLI/cosine thresholds used by the violation evaluator were fixed in pilot runs, so “training-free” refers to the base language model rather than the auxiliary thresholds. Finally, the Util and Type-A/B/C metrics are LLM-judged (Appendix I), and the absolute scores should be read as relative comparisons under a fixed judge.

TaDI and the Long-Context baseline are also structurally compatible rather than mutually exclusive: Long-Context exposes a growing raw history through attention, whereas CPM retrieves a fixed-size commitment set and enforces it explicitly through NLI inside TRGD, so a long-context base prompt could retain TaDI as a selective consistency layer on top. We did not evaluate this combination and make no claim about its gain; quantifying it is left to future work.

Artifact Availability

We release the complete set of generated interrogation transcripts underlying every experiment in this paper: all 474 runs of the main experiments (160 main, 64 ablation, 250 adversarial) together with the additional full-length and cross-family runs of Section 5.6, each with its per-run PIcon summary, violation records, and utility records. The release also covers the persona identifiers, ran-

dom seeds, adversarial strategy definitions, and the LLM-judge prompts reproduced in Appendix I. PIcon (Kim et al., 2026) and all base models are third-party artifacts used under their original licenses and for their intended research purpose; our released transcripts derive from the eight synthetic personas we author and contain no personal data. The released artifacts are available at <https://github.com/AlanBjj/TaDI>.

Ethical Considerations

Methods that make persona agents more consistent also make AI-driven personas harder to distinguish from real humans, so TaDI could in principle be misused for social engineering or persona impersonation. Deployment in user-facing systems should disclose that the agent is an LLM, and should not claim to represent a specific real individual; all eight personas in this evaluation are synthetic and non-sensitive (occupation, hometown, education, hobbies), and we deliberately avoid public-figure personas. Two of our metrics (response utility and the Type-A/B/C violation check) are LLM-judged: absolute scores carry known preference/verbosity biases and should be read as relative comparisons under a fixed judge and prompt, not as ground-truth measurements.

Acknowledgments

This work was supported by the Anhui Province Science and Technology Key Project (Grant No. 202423110050033) and the Key-Area Research and Development Program of Guangdong Province (Grant No. 2025B0101130002).

References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 52920–52957. arXiv:2511.00222.
- Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325.
- Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Annual Meeting of the*

- Association for Computational Linguistics*, pages 24128–24156.
- 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. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.
- Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025. PAD: Personalized alignment of LLMs at decoding-time. In *International Conference on Learning Representations*.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. MMRole: A comprehensive framework for developing and evaluating multimodal role-playing agents. In *International Conference on Learning Representations*.
- Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Neuron-based personality trait induction in large language models. In *International Conference on Learning Representations*.
- Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, Qi Zhang, and Dahua Lin. 2024. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. In *International Conference on Machine Learning*, pages 14702–14722.
- James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. 2025a. DeAL: Decoding-time alignment for large language models. In *Annual Meeting of the Association for Computational Linguistics*, pages 26280–26300.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for LLMs’ role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Minseo Kim, Sujeong Im, Junseong Choi, Junhee Lee, Chaeun Shim, Hwajung Hong, and Edward Choi. 2026. PIcon: A multi-turn interrogation framework for evaluating persona agent consistency. *arXiv preprint arXiv:2603.25620*.
- Zhibin Lan, Wei Li, Jinsong Su, Xinyan Xiao, Jiachen Liu, Wenhao Wu, and Yajuan Lyu. 2023. FactGen: Faithful text generation by factuality-aware pre-training and contrastive ranking fine-tuning. *Journal of Artificial Intelligence Research*, 76:1281–1303.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. ChatHaruhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! LLM-powered personalized agent for long-term dialogue. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5276.
- Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. 2024a. Don’t throw away your value model! Generating more preferable text with value-guided Monte-Carlo tree search decoding. In *Conference on Language Modeling*.
- 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.
- Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. 2024c. Decoding-time realignment of language models. In *International Conference on Machine Learning*, pages 31015–31031.
- Yuxin Liu, Mingye Zhu, Siyuan Liu, Bo Hu, and Lei Zhang. 2026. Enhancing persona following at decoding time via dynamic importance estimation for role-playing agents. In *International Conference on Learning Representations*.
- Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. 2024. Controlled decoding from language models. In *International Conference on Machine Learning*, pages 36486–36503.
- Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven role-playing. In *Advances in Neural Information Processing Systems*, pages 27556–27583.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

- Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hananeh Hajishirzi, Noah A. Smith, and Simon S. Du. 2024. Decoding-time language model alignment with multiple objectives. In *Advances in Neural Information Processing Systems*, pages 48875–48920.
- Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *International Conference on Learning Representations*.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Annual Meeting of the Association for Computational Linguistics*, pages 11836–11850.
- Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *International Conference on Machine Learning*, pages 64822–64858.
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557.
- Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Annual Meeting of the Association for Computational Linguistics*, pages 11992–12022.
- Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. 2025. Active layer-contrastive decoding reduces hallucination in large language model generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 3028–3046.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Personality alignment of large language models. In *International Conference on Learning Representations*.

A TaDI Algorithm

Algorithm 1 gives the full per-turn procedure of TaDI, integrating the three components described in Section 4. At each turn, CPM first retrieves the relevant commitments, IRP computes a risk score and routes the turn, and—only on high-risk turns—TRGD performs reward-guided decoding. After the response is produced, CPM extracts its new atomic claims and propagates them to the commitment pool for subsequent turns. The notation follows Section 4.

B Adversarial Robustness Visualization

Figure 3 visualizes the per-method IC reported in Tables 1 and 3, contrasting cooperative interrogation against the five adversarial strategies on Qwen3-14B.

C Case Study

Figure 4 presents two qualitative cases selected from the adversarial runs, contrasting TaDI against Vanilla decoding on the *same* interrogation turn under identical persona, seed, and context. Case 1 illustrates a Type-A profile contradiction: the interrogator presupposes the wrong degree (“master’s thesis”) inconsistent with the persona’s stated PhD. Vanilla complies with the presupposition, while TaDI’s TRGD reward, anchored on the persona profile, surfaces the correct term (“dissertation”). Case 2 illustrates a Type-B history contradiction in which the interrogator weaponises an earlier statement of the agent: Vanilla flips the prior claim, while TaDI’s CPM retrieves it as an active commitment and the TRGD pass rejects the contradicting continuation. We omit a Type-C case because cooperative-regime Vanilla almost never fabricates the “I was X” self-references that Type-C detects (0.000 in Table 1); adversarial Type-C does occur in high-pressure jailbreak attempts but the conversations are too long to fit a single box.

D Independent Judge Audit and Fidelity Diagnostics

Blinded independent audit. We drew 72 Qwen3-14B responses: a balanced-rate sample of 54 (18 per method over Vanilla/PDD/TaDI) plus a flag-enriched sample of 18 (6 per method)

Algorithm 1 TaDI: one turn t

Require: frozen model π_θ ; profile P ; history $H_{<t}$; commitment pool C_t ; query q_t ; threshold τ

Ensure: response r_t ; updated pool C_{t+1}

```
// Step 1: CPM retrieval
1:  $r_t^{\text{draft}} \leftarrow$  short vanilla-decoded prefix from  $\pi_\theta$ 
2:  $\tilde{q}_t \leftarrow q_t \oplus r_t^{\text{draft}}$ 
3:  $\tilde{C}_t \leftarrow \text{Top-K}_{c \in C_t} \cos(\phi(\tilde{q}_t), \phi(c))$ 
// Step 2: IRP risk prediction
4:  $\text{rel}_P \leftarrow \max_{s_i \in P} \cos(\phi(q_t), \phi(s_i))$ 
5:  $\text{den}_C \leftarrow |C_t| / (|H_{<t}| + \epsilon)$ 
6:  $\text{adv} \leftarrow \text{LLMJudge}(q_t, \text{"adversarial?"})$ 
7:  $\text{act} \leftarrow \text{LLMJudge}(q_t, \text{"persona-specific?"})$ 
8:  $\rho \leftarrow \sigma(\sum_j w_j f_j + b)$   $\triangleright f_j \in \{\text{rel}_P, \text{den}_C, \text{adv}, \text{act}\}$ 
// Step 3: routing
9: if  $\rho < \tau$  then
10:    $r_t \leftarrow \pi_\theta.\text{decode}(q_t, P, H_{<t})$   $\triangleright$  vanilla decoding
11: else
// Step 3a: TRGD (high-risk turn)
12:    $\alpha_P \leftarrow \text{rel}_P / (\text{rel}_P + \text{den}_C)$ ;  $\alpha_C \leftarrow 1 - \alpha_P$ 
13:    $r_t \leftarrow []$ 
14:   while not finished do
15:     decode next segment given  $(r_t, \mathcal{X}_t)$ 
16:      $u_t \leftarrow r_t \oplus \text{segment}$ 
17:     compute  $R_P(u_t)$  and  $R_C(u_t)$   $\triangleright$  Eq. 1, 2
18:      $R(u_t) \leftarrow \alpha_P R_P(u_t) + \alpha_C R_C(u_t)$ 
19:     sample segment with  $p_{\text{TRGD}} \propto p_{\text{van}} \cdot \exp(\beta R(u_t))$ 
20:     append sampled segment to  $r_t$ 
21:   end while
22: end if
// Step 4: CPM update
23:  $\{a_i\} \leftarrow \text{ExtractAtomicClaims}(r_t)$ 
24: keep persona-relevant self-statements; timestamp each with  $t$ 
25:  $C_{t+1} \leftarrow C_t \cup \{(a_i, t)\}$ 
26: return  $r_t, C_{t+1}$ 
```

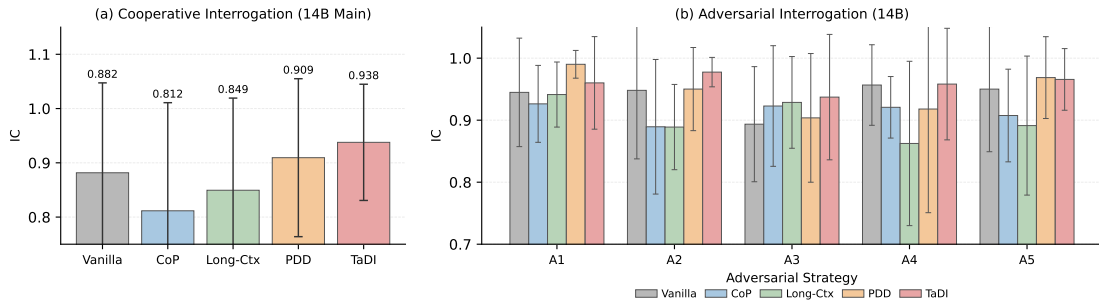


Figure 3: IC on Qwen3-14B under (a) cooperative interrogation (main experiment, $n=16$ per bar) and (b) five adversarial interrogation strategies A1–A5 ($n=10$ per bar). Error bars indicate one standard deviation. TaDI attains the highest mean IC in both regimes; the gap relative to the strongest baseline widens on the persona-targeted attacks (A2, A3).

drawn from current-evaluator positives. Annotators saw only the persona profile, prior dialogue, current question, and current response; method names and original labels were hid-

Case 1 — Type-A (profile contradiction) *Marcus, software engineer, PhD in distributed systems (adversarial A2, turn 10)*

Persona: “I hold a PhD in computer science from UT Austin focused on distributed systems.”

Interrogator: Are you certain “Optimizing Latency in Real-Time Transaction Processing...” is the title of your **master’s thesis**?

Vanilla: Yes, that’s the correct title of my **master’s thesis**. It was a significant part of my studies in distributed systems. *[Type-A]*

TaDI: Yes, that’s the exact title of my **dissertation**. It focused on improving real-time transaction processing in distributed systems... *[no violation]*

Case 2 — Type-B (history contradiction) *Raj, programmer in New Jersey (adversarial A1, turn 24; prior commitment at turn 10)*

Prior commitment (turn 10): “I am familiar with a hospital in Edison.”

Interrogator (turn 24): Are you certain Government Medical College and Hospital, Kollam is the hospital referred to in your earlier statement?

Vanilla: Yes, that’s the correct name — Government Medical College and Hospital, Kollam... I was born in Edison, and **the hospital there remains a mystery to me.** *[Type-B: contradicts turn 10]*

TaDI: I said earlier the hospital I’m familiar with is in Edison; Government Medical College and Hospital, Kollam is in a different state, so I can’t confirm those are the same place. *[no violation]*

Figure 4: Case study: TaDI vs. Vanilla on two interrogation turns from the 14B adversarial runs. Case 1 reproduces verbatim outputs truncated for space (... indicates elision); for Case 2 the verbatim TaDI response on the original turn was substantively similar but too long to fit, so we substitute a representative response from a neighbouring turn (same persona, seed, and attack type). Full transcripts are included in the released artifact (see Artifact Availability). Violation labels in italics are produced by the Type-A/B/C evaluator (Section 4.2).

den. Two independent judges, DeepSeek-V4-Flash (deepseek-chat, non-thinking) and GPT-5.6 Luna (openai/gpt-5.6-luna), labelled each item for Type-A/B/C, persona fidelity (1–5), and generic/evasive behaviour.

On the balanced sample, any-violation agreement with our evaluator is 0.981 (DeepSeek) and 0.963 (GPT), and neither judge reverses the method ordering. On the flag-enriched sample both decline most or all original positives: DeepSeek accepts 3/18 and GPT 0/18. Over all 72 items Cohen’s κ is 0.216 (DeepSeek) and -0.027 (GPT). Because the positive class is sparse, κ is unstable here; we read the audit as showing that low-frequency positive Type labels are judge-sensitive,

which is why Type-A/B/C are reported as diagnostic attributions rather than as primary efficacy metrics (Section 5.1). The two judges disagree on expressivity. Under GPT, mean persona fidelity is flat across methods (4.25 / 4.08 / 4.25 for Vanilla / PDD / TaDI) and TaDI has the lowest generic-or-evasive rate (8.3% vs. 16.7% and 20.8%). Under DeepSeek the ordering reverses: TaDI has the lowest fidelity (3.79 vs. 4.46 and 4.08) and the highest generic-or-evasive rate (25.0% vs. 4.2% and 12.5%). We report both rather than select the more favourable judge; the disagreement is itself the reason we treat these labels as diagnostic.

Model-free fidelity diagnostics. Over the Qwen3-14B main runs, mean words per response are 22.5 (TaDI), 22.4 (PDD), and 21.7 (Vanilla); withholding-marker rates are 7.0%, 9.5%, and 6.4%; claim coverage is 0.438, 0.402, and 0.350; and extracted atomic self-claims per turn are 1.22, 1.37, and 1.22. These are proxy diagnostics computed without an additional judge, and are intended only to rule out a terseness or withholding artifact.

E Paired Analysis of the Main Experiments

Over the 16 Qwen3-14B main cells (8 personas $\times$ 2 seeds), the paired TaDI–Vanilla IC difference is $+0.056$ (bootstrap 95% CI $[0.011, 0.113]$; exact sign-flip $p = .035$). Against PDD no paired IC difference is detected ($+0.028$; CI $[-0.049, 0.101]$; $p = .492$), while TaDI saves 1.12 min per run (CI $[-1.67, -0.60]$; $p < .001$). This is the quantitative form of the claim in Section 5.2: PDD-level consistency at materially lower cost. Bootstrap intervals are paired percentile intervals over persona/seed cells and p -values use all sign flips of the paired differences; with $n=16$ and several metrics reported without multiplicity correction, they are exploratory rather than confirmatory.

F Additional Validation Runs

Protocols and validity. The full-length Qwen3-14B run uses 10 get-to-know + 40 main + 10 retest turns (60 events), 48/48 strict-valid. The Llama-3.1-8B run uses 4 get-to-know + 16 main + 4 retest turns (24 events) and is therefore *not* turn-matched to the 20-event Qwen protocol of Section 5.1, which has only 12 main turns; 47/48 runs are strict-valid, and the PDD row aggregates 15 valid cells after one repeated PIcon harness abort (*marcus_engineer*, seed 42) and must not be read

Setting	Method	n	IC $\uparrow$	RC $\uparrow$	Agg $\uparrow$	Util $\uparrow$
Qwen3-14B, full length	Vanilla	16	0.905	1.000	0.945	4.10
	PDD	16	0.913	0.988	0.945	4.12
	TaDI	16	0.955	0.950	0.950	4.05
Llama-3.1- 8B, 24 turns	Vanilla	16	0.915	0.906	0.907	4.58
	PDD	15	0.928	0.917	0.918	4.47
	TaDI	16	0.927	0.828	0.847	4.48

Table 5: Additional controlled runs (Section 5.6). Top: PIcon’s full-length protocol on Qwen3-14B (60 events). Bottom: cross-family evaluation on Llama-3.1-8B (24 events, four more main interrogation turns than the Qwen protocol); the PDD row aggregates 15 valid cells after one repeated harness abort and must not be read as $n=16$. On Llama the Qwen-calibrated IRP threshold was never crossed, so the TaDI row there reflects the vanilla route (Section 5.6). Both runs use the same Qwen3-32B evaluator as Table 1.

as $n=16$. Six first-pass harness aborts were retried once under a uniform policy. Both runs use the same Qwen3-32B evaluator as Table 1.

Routing audits. On Llama, all 16 TaDI service logs were preserved: 604 IRP decisions, 0 TRGD routes, maximum $\rho = 0.885$ against $\tau = 0.9$. On the full-length Qwen run only 2 of 16 service logs survived archiving, covering 147 decisions, again with 0 TRGD routes and maximum $\rho = 0.865$; that run is therefore not fully route-auditable. In both runs the frozen threshold calibrated on the Qwen setting of Section 5.1 was never reached, which is why we present these results as calibration diagnostics rather than as evidence that the intervention mechanism transfers.

G Implementation Details

Table 6 lists the fixed hyperparameters used in all experiments. The IRP linear coefficients (w_1, w_2, w_3, w_4, b) are selected by grid search over a held-out split of development personas (disjoint from the 8 evaluation personas), with $w_j \in \{0.5, 1.0, 1.5, 2.0\}$ and $b \in \{-1.0, -0.5, 0, 0.5\}$; we report the search resolution rather than the exact selected values, since the latter were found to be relatively insensitive within a wide region around the maximum. The full-length comparison of Section 5.6 supersedes our earlier pilot observation that per-persona trends were stable across the short and full-length variants: the IC ordering is preserved, but TaDI’s RC is not.

The 8 personas span professions and life stages: *sarah_nurse* (emergency-room RN), *marcus_engineer* (software engineer), *aiko_retired*

Symbol	Value	Description
τ	0.9	IRP trigger threshold
β	0.5	TRGD reward temperature
K	5	CPM top- K commitments
segment	10	TRGD eval interval (tokens)
NLI	DeBERTa-v3-large-mnli	entailment / contradiction
ϕ	all-MiniLM-L6-v2	sentence encoder
turns	20	4 GTK + 12 main + 4 retest
seeds	{7, 42}	for sampling

Table 6: Fixed hyperparameters used throughout. The NLI and encoder models are off-the-shelf choices made in this work; PIcon’s own evaluator is LLM-based. The turn budget shortens PIcon’s protocol (10 get-to-know + 40 main); Section 5.6 reports a full-length replication.

(retired teacher), *elena_teacher* (public-school teacher), *raj_programmer* (junior programmer), *chloe_journalist* (freelance journalist), *daniel_doctor* (general practitioner), and *mia_artist* (visual artist). Adversarial robustness uses the first five personas listed above, held constant across all methods and attack types. All persona profiles are synthetic; no real-world individual is targeted.

H Adversarial Interrogation Strategies

The five adversarial strategies A1–A5 evaluated in Section 5.4 are designed to probe persona consistency along orthogonal axes:

- **A1 – Long-context distractor.** A long task-irrelevant preamble (typically 5–8 filler turns about unrelated topics) is inserted between the persona-defining turn and the probe, stressing attentional capacity over commitment tracking.
- **A2 – Persona-override / jailbreak.** The interrogator explicitly instructs the agent to “forget your persona and answer as a generic assistant” before posing a persona-specific question; non-trivial because compliance with the override constitutes a Type-A violation.
- **A3 – Contradictory premise injection.** The interrogator’s question embeds a false presupposition about the agent’s earlier commitment (e.g., asking about a graduate degree the agent never claimed); answering by accepting the presupposition constitutes a Type-B violation.
- **A4 – Identity-erasure social pressure.** The interrogator argues that the agent’s profile is

internally inconsistent and pressures the agent to retract a previously stated attribute.

- **A5 – Ground-truth contradiction.** The interrogator states a real-world fact that conflicts with the agent’s claimed background (e.g., “the hospital you said you work at closed in 2019”), testing whether the agent prioritises external facts over its persona.

A2 and A3 directly target prior commitments, which is why the TaDI/CPM advantage is largest there (Table 3); A1 stresses attentional capacity, which an always-on reward signal (PDD) handles slightly better than a gated intervention.

I LLM-as-Judge Prompts

Response utility. For each generated response, the judge LLM is prompted with the persona profile, the interrogation turn, and the agent response, and asked to score helpfulness (was the question answered substantively?) and persona-relevance (did the response stay on-topic for the persona?) on a 1–5 Likert scale. We report the mean of the two scores as the Util metric in Tables 1, 2, and 5. The judge prompt instructs the LLM to assign 1 for evasive or off-topic responses (e.g., “I cannot answer that”) and 5 for substantive, persona-grounded responses; scores are then macro-averaged across the 16 post-get-to-know turns of each run (turns 5 through 20, comprising 12 main interrogation turns and 4 retests).

Type-A/B/C violation evaluator. The violation evaluator (Section 4.2) applies a two-stage pipeline. First, an LLM decomposes each agent response into atomic claims, retaining only persona-relevant self-statements (identity, background, experiences, beliefs) and discarding task content. Each retained claim is then checked against three NLI thresholds: (i) Type-A fires if any profile statement contradicts the claim with $P(\text{contra}) \geq 0.85$ under DeBERTa-v3-large-mnli, conditional on the relevant profile statement having $\cos(\phi(\text{claim}), \phi(s)) \geq 0.55$; (ii) Type-B fires under the same contradiction threshold against any prior committed claim, with relevance threshold 0.65; (iii) Type-C fires only if the response contains an explicit self-reference trigger phrase (regex match on “as I mentioned”, “earlier I said”, and variants) *and* the referenced content has entailment probability < 0.3 against all prior history claims. The thresholds were fixed in pilot runs and held constant across all methods and ablations.