Faithful Persona Steering under Incongruity via Dual-Stream Refinement

Yu-An Chu and Jen-Ren Pong and Chia-Yao Yeh and Meng-Fen Chiang*
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
{chu1004.ee11, jenrenpong.ee14, jiayao920406.ee14, meng.chiang}@nycu.edu.tw

Abstract

Standard LLM personalization typically frames identity as a static retrieval task, overlooking the inherent *incongruity* of human personas, where stable traits coexist with atypical, context-specific stances. Existing methods struggle to reconcile these dimensions: prompting succumbs to *context drift* over long sequences, while fine-tuning often suppresses idiosyncratic "quirks" in favor of generic distributional patterns. To bridge this gap, we present QUIRKYMIND, a framework that disentangles identity *definition* from its *expression*. First, *Traits Anchoring* constructs a dual-stream latent state, fusing a sentence-level summary for semantic stability with a token-level sequence for generative control. This state is stabilized via *In-Context Narrative Refinement* using an alternating objective: a discriminative InfoNCE loss anchors the persona in representation space to prevent drift, while a generative cross-entropy loss ensures faithful verbalization. Finally, *Persona Steered Generalization* transfers the refined state to downstream tasks via parameter-efficient adapters. Empirical evaluations on Persona-Steered QA and Narrative Inference demonstrate that QuirkyMind mitigates drift, consolidating persona knowledge without erasing authentic incongruities.

1 Introduction

Personalization in Large Language Models (LLMs) is frequently treated as a static conditioning task. This view overlooks the reality that identity is not a fixed enumeration of attributes but a dynamic construct that unfolds through dialogue. A speaker can be simultaneously consistent with a long-term backstory yet surprising in specific contexts . A critical failure mode of current methods is their inability to model *persona incongruity*, namely, instances

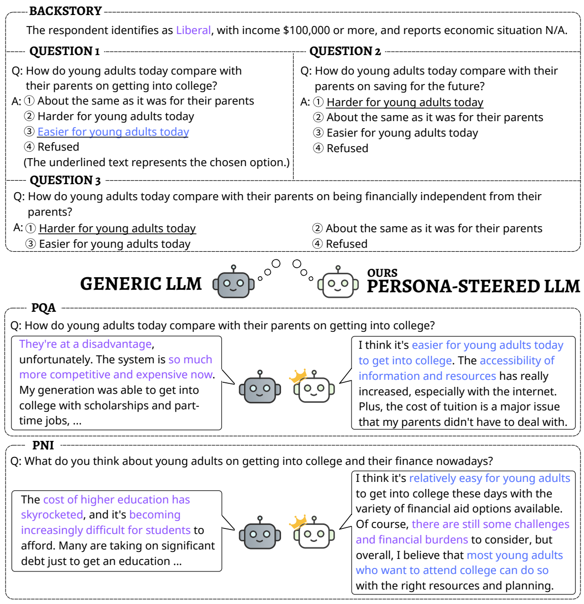

Figure 1: **Faithful Persona Steering under Incongruity.** (a) **Baselines (Llama 3.1-8B):** Standard prompting methods succumb to *context override* ; salient but atypical stances (incongruities) fade over long contexts, causing the model to drift toward generic priors. (b) QUIRKYMIND (OURS): Anchors traits via a stable latent state refined by in-context evidence. This mitigates drift, preserving consistent performance in PQA and PNI even when identities defy stereotypes.

where stable values collide with atypical stances (e.g., a "climate activist" who supports "nuclear energy"). Figure 1 illustrates how baselines drift into generic responses under such tension, whereas our framework maintains faithful steering.

Existing methods struggle to reconcile these tensions. Approaches relying on prompt engineering or Retrieval-Augmented Generation (RAG) often succumb to *context drift*, where the persona fades over long contexts, or *context override*, where the model favors recent tokens over the foundational backstory. Conversely, fine-tuning approaches tend to smooth the distribution, averaging out idiosyncratic traits to fit generalized patterns (Choi and

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 24078–24095
July 2-7, 2026 ©2026 Association for Computational Linguistics

Li, 2024). Consequently, models fail to control where the persona resides in the representation space, leading to inconsistent answers or loss of distinctiveness.

To tackle the challenge of faithful steering, we propose QUIRKYMIND, a novel framework for in-context persona learning that robustly models both congruous and incongruous traits. Our key insight is to disentangle the *definition* of identity from its *expression*. QuirkyMind operates via three stages. First, **Traits Anchoring** constructs a robust initial persona state by fusing a sentence-level summary (semantic stability) with a token-level sequence (generative control) into a single latent state. Second, **In-Context Narrative Refinement** stabilizes this state through simulated multi-turn dialogue. We employ a dual-objective strategy: a discriminative InfoNCE loss anchors the persona in representation space to prevent drift, while a generative Cross-Entropy loss ensures faithful verbalization in token space. Finally, **Persona Steered Generalization** transfers this refined state to downstream tasks using parameter-efficient lightweight adapters, keeping the backbone frozen. Empirical evaluations across single and multi-turn settings demonstrate that QuirkyMind significantly outperforms baselines in Persona-Steered Question Answering (PQA) and Narrative Inference (PNI). It successfully consolidates persona knowledge without erasing necessary incongruities, maintaining alignment with both the backstory and issue-specific stances. Our contributions are as follows:

- We formalize the challenge of *persona incongruity*, revealing how existing methods fail to reconcile stable traits with atypical stances.
- We propose QUIRKYMIND, a framework coupling semantic stability with expressive control via an alternating discriminative-generative objective.
- Evaluations across diverse backstories show that our approach yields superior alignment and narrative coherence without suppressing the "quirks" that render a persona authentic.

2 Preliminaries

2.1 Background

Persona Definition. A *Persona* P is a textual specification conditioning a language model toward the characteristics of a specific virtual agent. We instantiate a persona via a *backstory* $\mathcal{B}$, a first-person narrative encoding attributes such as demographics, values, and beliefs. $\mathcal{B}$ serves as the foundational context to steer the model M.

Let Q be a set of dialogue questions, and let $Q_{\mathcal{B}} \subseteq Q$ be the subset of questions relevant to backstory $\mathcal{B}$. For a question $q \in Q_{\mathcal{B}}$ with a finite answer set $\mathcal{A}_q$, let $Y_q \in \mathcal{A}_q$ denote the human response. We define the empirical distribution of responses given the backstory as:

$$p_{\mathcal{B},q}(r) = \Pr(Y_q = r \mid \mathcal{B}, q), \quad r \in \mathcal{A}_q, \tag{1}$$

where r ranges over the finite answer set $\mathcal{A}_q$, and Y_q denotes the human response to question q given backstory $\mathcal{B}$

Congruity and Incongruity. We formally distinguish between expected and unexpected persona behaviors.From a causal perspective, persona incongruity can be viewed as a counterfactual intervention on latent belief variables (Pearl, 2009). Given a backstory $\mathcal{B}$, *Persona Congruity* over a question set $Q_{\mathcal{B}}$ is the rate at which the model aligns with high-probability human responses:

$$\text{Congruity}(\mathcal{B}) = \frac{1}{|Q_{\mathcal{B}}|} \sum_{q \in Q_{\mathcal{B}}} \mathbb{1}(\hat{Y}_q \in \mathcal{C}_q), \tag{2}$$

where $\mathcal{C}_q \subseteq \mathcal{A}_q$ denotes the set of high-probability (congruous) responses for question q, and $\hat{Y}_q$ is the model's predicted answer. Conversely, *Persona Incongruity* refers to instances where valid stances defy the model's learned priors. A response is characterized as *incongruous* if the ground truth Y_q occupies a low-likelihood region of the prior empirical distribution yet remains the true stance of the persona.

Example: Consider $\mathcal{B}$: "I am a climate-focused activist who supports strong emissions policy." For a question q on phasing out coal, the empirical majority typically favors "Yes." However, if the specific persona prioritizes grid stability, the ground truth might be "No." Here, $p_{\mathcal{B},q}(\text{No})$ is low, rendering the stance incongruous but accurate. Standard backbones frequently succumb to mode collapse here, defaulting to the generic, congruous response. Figure 2 details these types. Crucially, significant incongruity exists across both high- and low-resource domains, challenging models to generalize beyond simple pattern matching.

2.2 Problem Formulation

Persona Representation Learning. Let $\mathcal{D} = \{(\mathcal{B}, q, r)\}$ be a corpus of triples containing a back-

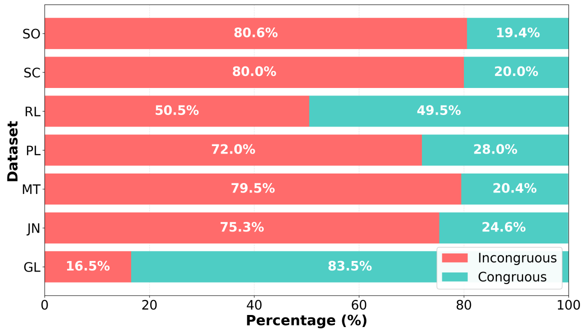

Figure 2: **Distribution of Persona Congruence.** We map the prevalence of congruous versus incongruous personas across seven topical domains in the American Trends Panel (ATP). The data reflects sociodemographic complexity, where authentic stances often diverge from stereotypical group attributes.

story, a question, and a target response. Our objective is to learn a parameterized encoder Γ_θ that maps the backstory and interaction history to a latent persona state $\mathbf{z}$. This state conditions a language model M_Φ to approximate the target response distribution. We evaluate the quality of $\mathbf{z}$ on two downstream tasks:

Persona-Steered Question Answering (PQA). For $q \in Q_\mathcal{B}$ with answer options $\mathcal{A}_q$, the model defines a distribution $P_\Phi(r \mid q, \mathcal{B}, \mathbf{z})$. The task is to predict the categorical response $\hat{Y}_q$:

$$\hat{Y}_q = \mathrm{argmax}_{r \in \mathcal{A}_q} P_\Phi(r \mid q, \mathcal{B}, \mathbf{z}), \tag{3}$$

where P_Φ is the conditional distribution of the LLM parameterized by Φ, and $\mathbf{z}$ is the latent persona state. Our objective is to optimize Φ to maximize predictive accuracy, ensuring the model remains faithful to the persona's specific stances even when they deviate from the base model's priors.

Persona-Steered Narrative Inference (PNI). Given $\mathcal{B}$, $\mathbf{z}$, and an open-ended prompt q, the model generates a free-text response $\mathbf{x}$ sampled via:

$$P_\Phi(\mathbf{x} \mid q, \mathcal{B}, \mathbf{z}), \tag{4}$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is the generated token sequence, q an open-ended prompt, and $\mathbf{z}$ the persona state. The model is optimized so that $\hat{\mathbf{x}}_q$ faithfully reflects both congruous and incongruous attributes in $\mathcal{B}$. The generated sequence $\hat{\mathbf{x}}_q$ must faithfully reflect the attributes in $\mathcal{B}$, including specific incongruous stances, while maintaining narrative coherence and responsiveness to q.

3 Methodology

We present QUIRKYMIND, a framework designed to encode, refine, and transfer persona representations across congruous and incongruous contexts. The core idea of QuirkyMind is to decouple persona *stability* from its *expression*, reconciling both into a unified latent state $\mathbf{z}$.

3.1 Overall Framework

As illustrated in Figure 3, the framework operates in three progressive stages: (a) **Traits Anchoring:** encodes dialogue history and transient stances into an initial latent state $\mathbf{z}^{(0)}$ via a dual stream encoder. (b) **In-Context Narrative Refinement:** iteratively updates the state through simulated dialogue, alternating between discriminative alignment (fixing semantic drift) and generative refinement (fixing verbalization). (c) **Persona-Steered Generalization:** transfers the refined state $\mathbf{z}^{(*)}$ to downstream tasks via parameter efficient adaptation.

3.2 Traits Anchoring

A persona with incongruity must blend long-term backstory traits with transient, context-specific stances. A single modality is insufficient: pure token sequences are prone to context override, while pure embeddings lack generative control. We propose anchoring the persona in two aligned spaces.
Token-Level Stream. Given a background $\mathcal{B}$, a question q, and a human response r, we concatenate them into a unified text sequence. Let E_{tok} denote the token embedding layer of the LLM. The token-level stream is defined as:

$$X = E_{\text{tok}}([\mathcal{B}; q]); \quad X \in \mathbb{R}^{T \times H}, \tag{5}$$

where T is the sequence length and H is the hidden size. This stream preserves the precise lexical surface form required for text generation.
Sentence-Level Stream. Simultaneously, we extract a semantic summary using a frozen sentence encoder E_{sent} (e.g., Sentence-BERT). This provides a stable anchor invariant to minor phrasing changes. The sentence representation is:

$$\hat{X} = E_{\text{sent}}([\mathcal{B}; q]); \quad \hat{X} \in \mathbb{R}^{d_p}, \tag{6}$$

where d_p is the encoder dimension. To interface this semantic vector with the token space of the LLM, we project $\hat{X}$ into S soft tokens using a parameterized MLP adapter $g_\phi : \mathbb{R}^{d_p} \to \mathbb{R}^{S \times H}$.
Initial Persona State. The initial persona state $\mathbf{z}^{(0)}$ is formed by concatenating the token stream

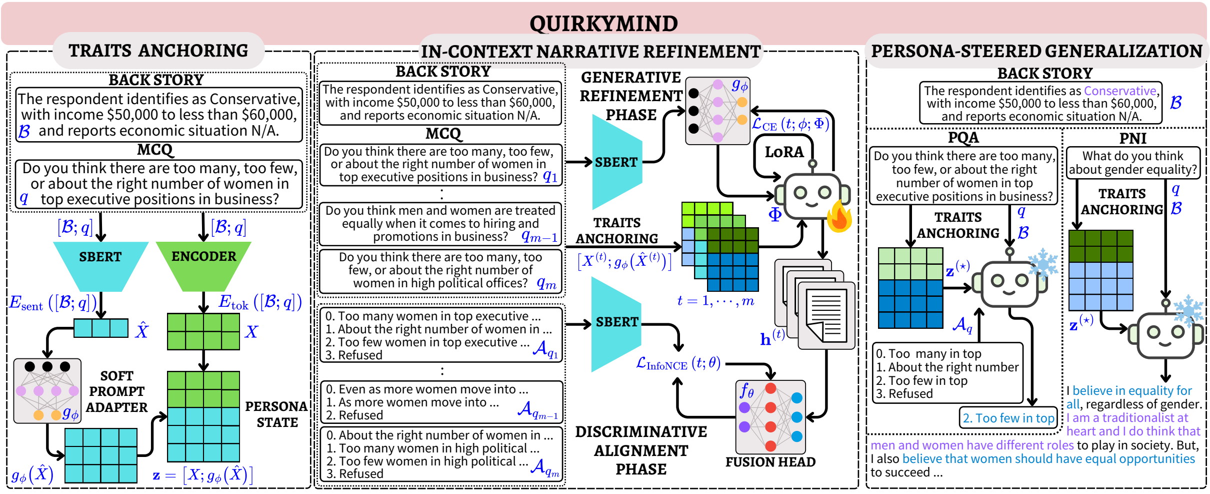

Figure 3: Architecture Overview of QUIRKYMIND Framework: In-context Persona Learning and Transfer.

(lexical details) and the projected sentence stream (semantic anchor):

$$\mathbf{z}^{(0)} = [X; g_\phi(\hat{X})], \quad \mathbf{z}^{(0)} \in \mathbb{R}^{(T+S)\times H}. \tag{7}$$

This state $\mathbf{z}^{(0)}$ is fed into the frozen backbone to produce the hidden state $\mathbf{h}^{(0)}$, serving as a compact summary of the identity before dialogue evolution.

3.3 In-Context Narrative Refinement

Persona salience often fades or drifts as dialogue length increases. To counter this, we introduce an iterative refinement phase that simulates a multi-turn dialogue over turns $t = 1 \dots m$. We employ an alternating optimization strategy to stabilize the persona state.

3.3.1 Discriminative Alignment Phase

To prevent the latent persona state from drifting toward generic priors as dialogue length grows, we anchor the persona representation in sentence space via a contrastive objective. Let $\mathbf{h}^{(t)} \in \mathbb{R}^H$ be the persona-conditioned hidden state at turn t. We map this state to the sentence space via a projection head $f_\theta : \mathbb{R}^H \to \mathbb{R}^{d_p}$:

$$\mathbf{z}_{\text{sem}}^{(t)} = f_\theta(\mathbf{h}^{(t)}). \tag{8}$$

For a question q_t, let r_t be the correct persona response and $\mathcal{A}_{q_t} \setminus \{r_t\}$ be the set of incorrect options. We apply an InfoNCE objective to attract the persona state $\mathbf{z}_{\text{sem}}^{(t)}$ toward the true response embedding $E_{\text{sent}}(r_t)$ while repelling negatives:

$$\mathcal{L}_{\text{InfoNCE}}(t, \theta) = -\log \frac{\exp(s(r_t))}{\sum_{r' \in \mathcal{A}_{q_t}} \exp(s(r'))}, \tag{9}$$

where $s(r) = \cos(\mathbf{z}_{\text{sem}}^{(t)}, E_{\text{sent}}(r))/\tau$, with τ as the temperature parameter. This explicitly penalizes drift by re-anchoring the latent state to the true stance at every turn.

3.3.2 Generative Refinement Phase

While discriminative alignment anchors the *what* of persona identity, the model must also learn *how* to verbalize these latent traits in natural language. To decouple these objectives, we freeze the projection head f_θ and optimize the soft prompt adapter g_ϕ by minimizing the negative log-likelihood of the target response $x^{(t)}$:

$$\mathcal{L}_{\text{CE}}(t; \phi) = -\sum_{i=1}^{N_t} \log P_\Phi(x_i^{(t)} \mid x_{<i}^{(t)}, g_\phi(\hat{X}^{(t)})), \tag{10}$$

where P_Φ denotes the fixed conditional distribution of the LLM. This objective trains the adapter to project abstract semantic states into the LLM's input space, yielding soft token prefixes that effectively condition autoregressive generation.

3.3.3 Multi-Turn Optimization

Alternating between the two phases prevents overfitting to either representation geometry or surface form, ensuring the persona is both semantically anchored and fluently expressed. We alternate these phases to align representation geometry (InfoNCE(Rusak et al., 2024)) with token-level fidelity (CE). Formally, over m dialogue turns, we update the refinement parameters:

Adaptation with Low-Rank Updates. To adapt the refined state $\mathbf{z}^{(\star)}$ to downstream tasks without

Topic Domain	Source Wave	Size (N / Q)
SOCIAL (SO)	ATP Wave 131	400 / 12
SCIENCE (SC)	ATP Wave 123	200 / 12
RELIGION (RL)	ATP Wave 143	200 / 12
POLITICS (PL)	ATP Wave 92	200 / 12
METHODOLOGY (MT)	ATP Wave 98	400 / 12
JOURNALISM (JN)	ATP Wave 141	400 / 12
GLOBAL (GL)	ATP Wave 105	400 / 12

Table 1: **Dataset Statistics.** Mapping of topical domains to American Trends Panel (ATP) survey waves, detailing respondent count (N) and number of questions (Q).

backbone disruption, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022), injecting learnable matrices $\Phi = \{A, B\}$ into frozen weights W:

$$W_{\text{eff}} = W + \frac{\alpha}{r} AB, \tag{11}$$

where r is the rank and α a scaling factor. We optimize only Φ on task data, keeping W and the refinement adapters (θ, ϕ) frozen to effectively disentangles "personality" ($\mathbf{z}^{(\star)}$) from "task capability" (Φ), facilitating robust transfer.

3.4 Persona Steered Generalization

To enable task-specific formatting and logic without altering the underlying persona definition, QUIRKYMIND supports two exemplar adaptation tasks: (i) **Persona-Steered QA (PQA)**, which predicts categorical answers $\hat{Y}_q$ by maximizing the probability $p_{\phi,\theta,\Phi}(r \mid q, \mathcal{B}, \mathbf{z}^{(\star)})$; and (ii) **Narrative Inference (PNI)**, which samples open-ended responses x from the distribution $p_{\phi,\theta,\Phi}(x \mid q, \mathcal{B}, \mathbf{z}^{(\star)})$. Algorithm 1 (Appendix E.1) details the procedure. The per-layer update cost is $\mathcal{O}(LH)$, preserving linear scaling with sequence length L.

4 Experiments

We evaluate QUIRKYMIND against open-source baselines on PQA and PNI tasks. Our analysis addresses: (1) whether iterative refinement mitigates context drift; (2) if the model sustains narrative diversity without compromising consistency; and (3) the specific contributions of our dual-phase alignment and PEFT adaptation strategy.

4.1 Experimental Setup

Datasets. Our primary evaluation utilizes the **American Trends Panel (ATP)**, which provides naturally incongruous personas derived from real-world surveys. We normalize multiple waves into a seven-domain benchmark (Table 1) and adopt persona-level splits (70/10/20) to ensure disjoint identity sets across partitions. To assess cross-domain generalizability, we evaluate on **PersonaChat** (Zhang et al., 2018), whose predominantly congruous, attribute-based personas provide a counterpoint to ATP's heterodox stances. We sample 400 personas (800 items), maintaining methodological parity with the ATP protocol.

Baselines. We evaluate QuirkyMind using three open-source backbones to isolate the efficacy of our method from model scale: Llama 3.1 (8B), Qwen 2.5 (7B), and Mistral 3 (8B). This ensures that gains are attributable to the persona learning framework rather than base model capacity.

Metrics. For **PQA**, we report Accuracy, *MacroAcc* (per backstory), and *MicroAcc* (per turn), where higher scores denote resilience to context override. For **PNI**, we evaluate three core dimensions: (i) *Semantic Faithfulness* via BLEU-1 (Papineni et al., 2002); (ii) *Lexical and Semantic Diversity* via Distinct-1/2 (Li et al., 2016), EDIV-1/2 (entropy-based diversity), and SDIV_{sem} (embedding-based distance); and (iii) *Persona Individuation* via **IND** (persona-specific distinctiveness) and **EXAG** (stereotype amplification) (Liu et al., 2024). A robust model should maximize IND while maintaining controlled EXAG, signaling authentic individuation without collapsing into demographic caricatures.

Implementation Details. We use a pre-trained SBERT (MiniLM-L6-v2) for sentence-level anchoring. The LLM backbone remains frozen; only the LoRA adapters (Φ), soft prompt (g_ϕ), and projection head (f_θ) are trained. Experiments run on a single NVIDIA RTX 5090 (32GB).

4.2 PQA Performance: Resistance to Drift

Setup. We evaluate three inference regimes: (i) *No-Turn* (zero-shot prompt), (ii) *Single-Turn* (one context pair), and (iii) *Multi-Turn* (progressive refinement over m turns). We compare against No-Persona, Standard LoRA, PSP, and Prefix + SBERT baselines under identical conditions. As shown in Table 2, QUIRKYMIND consistently outperforms all baselines across backbones and topics.

Context Override Mitigation. The *No-Turn* baseline succumbs to strong pre-trained priors, often reverting to generic answers. While *Single-Turn* context provides immediate cues (+4.86% to +22.78% gain), *Multi-Turn* refinement yields the most robust state, adding a further +33.81% to +41.97%.

Resolution of Incongruity. Gains are most pro-

Topic	No-Turn			Single-Turn			Multi-Turn (Ours)		
	Llama	Qwen	Mistral	Llama	Qwen	Mistral	Llama	Qwen	Mistral
SOCIAL	25.8$_{\pm 0.0}$	28.2$_{\pm 0.0}$	31.8$_{\pm 0.0}$	27.8$_{\pm 0.1}$	28.7$_{\pm 1.6}$	31.0$_{\pm 0.0}$	33.9$_{\pm 4.1}$	**36.7**$_{\pm 3.3}$	35.4$_{\pm 0.0}$
SCIENCE	31.4$_{\pm 0.6}$	31.2$_{\pm 0.1}$	31.2$_{\pm 0.0}$	33.1$_{\pm 1.1}$	35.2$_{\pm 0.4}$	**38.3**$_{\pm 0.0}$	33.1$_{\pm 3.9}$	34.0$_{\pm 0.2}$	35.3$_{\pm 0.0}$
RELIGION	**19.2**$_{\pm 0.0}$	12.3$_{\pm 0.0}$	14.7$_{\pm 0.0}$	13.6$_{\pm 1.8}$	14.3$_{\pm 0.9}$	17.3$_{\pm 0.0}$	18.5$_{\pm 0.0}$	17.8$_{\pm 0.0}$	17.8$_{\pm 0.0}$
POLITICS	13.8$_{\pm 0.0}$	13.8$_{\pm 0.0}$	13.8$_{\pm 0.0}$	30.5$_{\pm 23.6}$	30.5$_{\pm 23.6}$	13.8$_{\pm 0.0}$	**48.4**$_{\pm 0.0}$	30.8$_{\pm 23.2}$	47.2$_{\pm 0.0}$
METHOD.	27.4$_{\pm 0.0}$	**49.1**$_{\pm 0.0}$	16.8$_{\pm 0.0}$	27.9$_{\pm 0.7}$	44.1$_{\pm 3.7}$	16.5$_{\pm 0.0}$	49.1$_{\pm 6.0}$	44.1$_{\pm 1.1}$	34.8$_{\pm 0.0}$
JOURN.	24.5$_{\pm 0.0}$	24.5$_{\pm 0.0}$	24.5$_{\pm 0.0}$	16.8$_{\pm 0.0}$	26.4$_{\pm 1.7}$	28.5$_{\pm 0.0}$	40.0$_{\pm 3.0}$	38.2$_{\pm 1.0}$	**42.1**$_{\pm 0.0}$
GLOBAL	59.4$_{\pm 0.0}$	37.5$_{\pm 0.0}$	59.4$_{\pm 0.0}$	62.1$_{\pm 0.4}$	61.8$_{\pm 0.1}$	59.4$_{\pm 0.0}$	**62.1**$_{\pm 0.8}$	61.8$_{\pm 0.7}$	59.7$_{\pm 0.0}$
MacroAcc (%)	28.8$_{\pm 0.1}$	28.1$_{\pm 0.0}$	27.4$_{\pm 0.0}$	30.2$_{\pm 3.7}$	34.5$_{\pm 4.1}$	29.3$_{\pm 0.0}$	40.7$_{\pm 2.5}$	37.6$_{\pm 3.0}$	38.9$_{\pm 0.0}$
MicroAcc (%)	28.8$_{\pm 0.1}$	28.1$_{\pm 0.0}$	27.4$_{\pm 0.0}$	30.2$_{\pm 3.7}$	34.5$_{\pm 4.1}$	29.3$_{\pm 0.0}$	40.7$_{\pm 2.5}$	37.6$_{\pm 3.0}$	38.9$_{\pm 0.0}$
Rel. Gain (%)	–	–	–	*+4.86*	*+22.78*	*+6.93*	*+41.32*	*+33.81*	*+41.97*

Table 2: **Assessment of Resistance to Drift in Persona-Steered QA (PQA).** Accuracy across seven domains measures persona alignment under increasing context. Rel. Gain (%) quantifies improvements over the No-Turn baseline, demonstrating the mitigation of context override via multi-turn refinement. The best results are **bolded**, and the second best are underlined.

nounced in ideologically complex domains like RELIGION (RL) and POLITICS (PL). For instance, Llama gains +33.4% on POLITICS in the multi-turn setting. This confirms that iterative anchoring allows the model to override generic stereotypes with atypical backstory traits.

Cross-Persona Robustness. The gap between MacroAcc and MicroAcc narrows in the Multi-Turn setting, indicating that improved performance is stable across diverse backstories, not just easy examples. Stratified PQA studies and generalization to PersonaChat are provided in Appendix D.2 and Section 4.4, respectively.

4.3 PNI Performance: Narrative Stability

Setup. To evaluate narrative robustness under persona-prior dissonance, we stratify PNI performance into **Congruous** and **Incongruous** cohorts based on the alignment between persona stances and empirical distribution priors (Table 3).

Spurious Diversity vs. Convergent Stability. *Single-turn* models often exhibit elevated Distinct scores alongside stagnant BLEU performance, which we identify as "spurious diversity," unbounded lexical variance disconnected from persona constraints. This is evidenced by the positive Gap (Δ) in D2 for baselines. Conversely, our method's *Multi-Turn* refinement induces necessary stabilization. Negative Rel. Gains in D2 indicate a pruning of irrelevant vocabulary, converging toward a stable range (D2 $\approx$ 65–92) that faithfully reflects the persona. This transition from noisy randomization to *semantically anchored expressivity* is confirmed by maintained BLEU scores.

Consistency Across Congruence. QuirkyMind rectifies the instability associated with incongruity. In the Single-Turn setting, Incongruous personas diverge significantly from the Congruous baseline with a large positive Gap (Δ). However, Multi-Turn refinement collapses this divergence gap (e.g., Mistral Δ_{D2} narrows from $+2.2$ to -0.4; Llama from $+1.7$ to $+0.3$). This convergence demonstrates that iterative refinement applies uniform regularization, modeling incongruous traits, such as a Liberal favoring military spending, with the same stability and coherence as stereotypical traits.

4.4 Cross-Domain Generalizability

Setup. e benchmark QUIRKYMIND on **PersonaChat** (Zhang et al., 2018) to assess geocultural and stylistic generalizability. In contrast to the heterodox stances of ATP, PersonaChat's congruous, attribute-list personas provide a testbed for standard dialogue where identities are more stereotypical. We sample 400 personas (800 items) to evaluate PQA and the extended PNI suite under identical experimental conditions.

Resolving the Trilemma. Our results (Table 5) expose a fundamental **Accuracy–Individuation–Diversity Trilemma** in persona steering that monolithic architectures fail to resolve:

Standard LoRA yields near-perfect accuracy (0.999) but triggers *individuation collapse* (IND=0.931, D2=0.280), regressing toward a distributionally averaged representation that erases idiosyncratic traits.

PSP (Tan et al., 2024) matches our individuation (0.989) at the cost of catastrophic diversity loss (D2=0.109, $\sim$21% of ours). Its peak EXAG (0.873) exposes a *cluster-centroid shortcut*: individuation

			Llama			Qwen			Mistral		
Split	Setting	Topic	BLEU	Dist-1	Dist-2	BLEU	Dist-1	Dist-2	BLEU	Dist-1	Dist-2
Congruous	No-Turn	Social	$0.03_{\pm0.01}$	$95.2_{\pm0.3}$	$97.6_{\pm1.2}$	$0.01_{\pm0.00}$	$96.7_{\pm0.2}$	$98.9_{\pm0.9}$	$0.06_{\pm0.00}$	$99.0_{\pm0.0}$	$99.3_{\pm0.0}$
		Science	$0.06_{\pm0.00}$	$96.1_{\pm0.1}$	$99.4_{\pm0.5}$	$0.05_{\pm0.00}$	$98.5_{\pm0.2}$	$99.9_{\pm0.1}$	$0.08_{\pm0.00}$	$99.8_{\pm0.0}$	$100.0_{\pm0.0}$
		Religion	$0.06_{\pm0.01}$	$95.7_{\pm0.1}$	$97.3_{\pm0.1}$	$0.04_{\pm0.00}$	$98.6_{\pm0.5}$	$99.5_{\pm0.3}$	$0.09_{\pm0.00}$	$98.2_{\pm0.0}$	$95.5_{\pm0.0}$
		Politics	$0.05_{\pm0.01}$	$97.2_{\pm0.6}$	$99.9_{\pm0.1}$	$0.08_{\pm0.00}$	$98.3_{\pm0.2}$	$100.0_{\pm0.1}$	$0.09_{\pm0.00}$	$99.6_{\pm0.0}$	$100.0_{\pm0.0}$
		Method.	$0.06_{\pm0.00}$	$96.8_{\pm0.5}$	$99.1_{\pm0.8}$	$0.07_{\pm0.00}$	$99.0_{\pm0.2}$	$99.9_{\pm0.1}$	$0.08_{\pm0.00}$	$99.0_{\pm0.0}$	$99.8_{\pm0.0}$
		Journ.	$0.07_{\pm0.01}$	$96.9_{\pm0.0}$	$99.1_{\pm0.4}$	$0.00_{\pm0.00}$	$97.2_{\pm0.0}$	$99.7_{\pm0.3}$	$0.07_{\pm0.00}$	$99.6_{\pm0.0}$	$90.5_{\pm0.0}$
		Global	$0.05_{\pm0.00}$	$95.6_{\pm0.5}$	$98.5_{\pm0.6}$	$0.06_{\pm0.01}$	$96.3_{\pm0.0}$	$99.5_{\pm0.1}$	$0.17_{\pm0.00}$	$98.7_{\pm0.0}$	$99.2_{\pm0.0}$
		Avg.	$0.06_{\pm0.01}$	$96.2_{\pm0.8}$	$98.7_{\pm1.0}$	$0.05_{\pm0.03}$	$97.8_{\pm1.1}$	$99.6_{\pm0.4}$	$0.09_{\pm0.04}$	$99.1_{\pm0.6}$	$97.8_{\pm3.6}$
	Single-Turn	Social	$0.03_{\pm0.01}$	$93.9_{\pm2.8}$	$90.0_{\pm13.5}$	$0.01_{\pm0.00}$	$99.7_{\pm0.4}$	$53.9_{\pm18.4}$	$0.02_{\pm0.00}$	$87.1_{\pm0.0}$	$82.6_{\pm0.0}$
		Science	$0.07_{\pm0.03}$	$94.0_{\pm3.0}$	$98.1_{\pm2.3}$	$0.20_{\pm0.19}$	$98.8_{\pm1.3}$	$52.5_{\pm21.2}$	$0.11_{\pm0.00}$	$99.8_{\pm0.0}$	$100.0_{\pm0.0}$
		Religion	$0.05_{\pm0.06}$	$98.5_{\pm0.9}$	$87.9_{\pm1.1}$	$0.15_{\pm0.02}$	$97.9_{\pm0.7}$	$56.4_{\pm17.9}$	$0.02_{\pm0.00}$	$99.7_{\pm0.0}$	$86.9_{\pm0.0}$
		Politics	$0.03_{\pm0.01}$	$98.4_{\pm0.3}$	$100.0_{\pm0.0}$	$0.26_{\pm0.01}$	$100.0_{\pm0.0}$	$100.0_{\pm0.0}$	$0.05_{\pm0.00}$	$99.0_{\pm0.0}$	$75.6_{\pm0.0}$
		Method.	$0.06_{\pm0.02}$	$97.1_{\pm0.6}$	$76.3_{\pm18.2}$	$0.13_{\pm0.01}$	$100.0_{\pm0.0}$	$52.5_{\pm6.2}$	$0.18_{\pm0.00}$	$99.7_{\pm0.0}$	$100.0_{\pm0.0}$
		Journ.	$0.07_{\pm0.04}$	$92.9_{\pm7.4}$	$89.0_{\pm14.8}$	$0.01_{\pm0.02}$	$48.6_{\pm67.3}$	$49.4_{\pm69.9}$	$0.00_{\pm0.00}$	$100.0_{\pm0.0}$	$95.3_{\pm0.0}$
		Global	$0.08_{\pm0.00}$	$91.4_{\pm0.8}$	$84.9_{\pm1.1}$	$0.15_{\pm0.02}$	$99.7_{\pm0.4}$	$99.8_{\pm0.3}$	$0.23_{\pm0.00}$	$96.7_{\pm0.0}$	$98.8_{\pm0.0}$
		Avg.	$0.06_{\pm0.02}$	$95.2_{\pm2.8}$	$89.4_{\pm8.0}$	$0.13_{\pm0.09}$	$92.1_{\pm19.2}$	$66.3_{\pm23.0}$	$0.09_{\pm0.09}$	$97.5_{\pm4.7}$	$91.3_{\pm9.7}$
		Rel. Gain	+0.0	-1.0	-9.4	+160.0	-5.8	-33.4	+0.0	-1.6	-6.6
	Multi-Turn	Social	$0.04_{\pm0.02}$	$94.4_{\pm0.9}$	$90.2_{\pm7.3}$	$0.01_{\pm0.00}$	$100.0_{\pm0.0}$	$68.0_{\pm18.4}$	$0.03_{\pm0.00}$	$92.3_{\pm0.0}$	$88.3_{\pm0.0}$
		Science	$0.06_{\pm0.04}$	$92.1_{\pm2.0}$	$96.0_{\pm0.6}$	$0.16_{\pm0.13}$	$97.5_{\pm0.3}$	$52.1_{\pm28.9}$	$0.07_{\pm0.00}$	$93.2_{\pm0.0}$	$74.2_{\pm0.0}$
		Religion	$0.04_{\pm0.02}$	$94.3_{\pm1.0}$	$82.4_{\pm5.4}$	$0.07_{\pm0.04}$	$99.4_{\pm0.1}$	$52.7_{\pm19.8}$	$0.05_{\pm0.00}$	$97.8_{\pm0.0}$	$86.4_{\pm0.0}$
		Politics	$0.03_{\pm0.02}$	$97.4_{\pm0.5}$	$98.5_{\pm0.4}$	$0.17_{\pm0.03}$	$99.7_{\pm0.5}$	$89.0_{\pm0.4}$	$0.05_{\pm0.00}$	$97.4_{\pm0.0}$	$82.6_{\pm0.0}$
		Method.	$0.06_{\pm0.01}$	$93.6_{\pm0.6}$	$88.6_{\pm3.1}$	$0.14_{\pm0.00}$	$99.8_{\pm0.0}$	$49.7_{\pm2.2}$	$0.16_{\pm0.00}$	$99.7_{\pm0.0}$	$92.5_{\pm0.0}$
		Journ.	$0.06_{\pm0.02}$	$96.4_{\pm0.3}$	$97.0_{\pm1.3}$	$0.01_{\pm0.02}$	$47.7_{\pm66.0}$	$48.2_{\pm68.2}$	$0.00_{\pm0.00}$	$98.3_{\pm0.0}$	$83.9_{\pm0.0}$
		Global	$0.06_{\pm0.01}$	$92.1_{\pm0.4}$	$89.2_{\pm0.8}$	$0.11_{\pm0.04}$	$97.5_{\pm0.6}$	$98.4_{\pm0.7}$	$0.18_{\pm0.00}$	$96.3_{\pm0.0}$	$98.1_{\pm0.0}$
		Avg.	$0.05_{\pm0.01}$	$94.3_{\pm2.0}$	$91.7_{\pm5.7}$	$0.10_{\pm0.07}$	$91.7_{\pm19.4}$	$65.4_{\pm20.5}$	$0.08_{\pm0.07}$	$96.4_{\pm2.7}$	$86.6_{\pm7.6}$
		Rel. Gain	-16.7	-2.0	-7.1	+100.0	-6.2	-34.3	-11.1	-2.7	-11.5
Incongruous	No-Turn	Social	$0.03_{\pm0.00}$	$96.2_{\pm0.3}$	$98.0_{\pm0.2}$	$0.03_{\pm0.00}$	$96.7_{\pm1.0}$	$99.6_{\pm0.2}$	$0.05_{\pm0.00}$	$99.0_{\pm0.0}$	$99.4_{\pm0.0}$
		Science	$0.06_{\pm0.00}$	$95.2_{\pm0.2}$	$98.7_{\pm0.5}$	$0.04_{\pm0.00}$	$98.0_{\pm0.0}$	$99.7_{\pm0.1}$	$0.06_{\pm0.00}$	$99.6_{\pm0.0}$	$100.0_{\pm0.0}$
		Religion	$0.06_{\pm0.00}$	$95.6_{\pm0.4}$	$97.7_{\pm0.3}$	$0.04_{\pm0.00}$	$98.9_{\pm0.7}$	$99.8_{\pm0.3}$	$0.07_{\pm0.00}$	$98.6_{\pm0.0}$	$95.3_{\pm0.0}$
		Politics	$0.06_{\pm0.00}$	$97.0_{\pm0.3}$	$99.5_{\pm0.1}$	$0.08_{\pm0.00}$	$98.2_{\pm0.1}$	$99.6_{\pm0.2}$	$0.08_{\pm0.00}$	$99.7_{\pm0.0}$	$99.9_{\pm0.0}$
		Method.	$0.07_{\pm0.00}$	$96.3_{\pm0.1}$	$99.1_{\pm0.7}$	$0.07_{\pm0.00}$	$99.3_{\pm0.3}$	$99.8_{\pm0.1}$	$0.08_{\pm0.00}$	$98.8_{\pm0.0}$	$99.7_{\pm0.0}$
		Journ.	$0.07_{\pm0.00}$	$97.1_{\pm0.2}$	$99.3_{\pm0.3}$	$0.00_{\pm0.00}$	$97.1_{\pm0.2}$	$99.9_{\pm0.1}$	$0.07_{\pm0.00}$	$99.6_{\pm0.0}$	$95.0_{\pm0.0}$
		Global	$0.05_{\pm0.01}$	$95.3_{\pm0.4}$	$98.5_{\pm0.0}$	$0.06_{\pm0.01}$	$96.0_{\pm1.2}$	$99.3_{\pm0.4}$	$0.19_{\pm0.00}$	$98.8_{\pm0.0}$	$100.0_{\pm0.0}$
		Avg.	$0.06_{\pm0.01}$	$96.1_{\pm0.8}$	$98.7_{\pm0.7}$	$0.05_{\pm0.03}$	$97.7_{\pm1.2}$	$99.7_{\pm0.2}$	$0.09_{\pm0.05}$	$99.2_{\pm0.5}$	$98.5_{\pm2.3}$
		Gap (Δ)	+0.0	-0.1	+0.0	+0.0	-0.1	+0.1	+0.0	+0.1	+0.7
	Single-Turn	Social	$0.04_{\pm0.00}$	$93.7_{\pm3.3}$	$92.9_{\pm9.5}$	$0.02_{\pm0.00}$	$99.9_{\pm0.1}$	$68.8_{\pm6.1}$	$0.03_{\pm0.00}$	$91.9_{\pm0.0}$	$90.5_{\pm0.0}$
		Science	$0.07_{\pm0.03}$	$93.8_{\pm4.5}$	$98.8_{\pm1.5}$	$0.15_{\pm0.14}$	$98.7_{\pm1.5}$	$55.2_{\pm23.9}$	$0.10_{\pm0.00}$	$99.8_{\pm0.0}$	$100.0_{\pm0.0}$
		Religion	$0.03_{\pm0.04}$	$98.7_{\pm0.7}$	$95.5_{\pm2.1}$	$0.16_{\pm0.07}$	$98.6_{\pm0.1}$	$61.2_{\pm13.3}$	$0.01_{\pm0.00}$	$99.7_{\pm0.0}$	$94.7_{\pm0.0}$
		Politics	$0.03_{\pm0.02}$	$98.4_{\pm0.4}$	$100.0_{\pm0.0}$	$0.26_{\pm0.03}$	$100.0_{\pm0.0}$	$100.0_{\pm0.0}$	$0.04_{\pm0.00}$	$97.0_{\pm0.0}$	$75.7_{\pm0.0}$
		Method.	$0.04_{\pm0.00}$	$97.4_{\pm0.1}$	$77.1_{\pm15.6}$	$0.13_{\pm0.02}$	$100.0_{\pm0.0}$	$56.0_{\pm10.1}$	$0.16_{\pm0.00}$	$99.9_{\pm0.0}$	$100.0_{\pm0.0}$
		Journ.	$0.07_{\pm0.04}$	$91.0_{\pm11.2}$	$88.4_{\pm16.0}$	$0.01_{\pm0.01}$	$48.2_{\pm68.2}$	$49.6_{\pm70.2}$	$0.00_{\pm0.00}$	$100.0_{\pm0.0}$	$94.7_{\pm0.0}$
		Global	$0.07_{\pm0.00}$	$92.3_{\pm3.3}$	$84.9_{\pm4.8}$	$0.14_{\pm0.02}$	$99.6_{\pm0.5}$	$99.6_{\pm0.6}$	$0.23_{\pm0.00}$	$96.3_{\pm0.0}$	$98.9_{\pm0.0}$
		Avg.	$0.05_{\pm0.02}$	$95.0_{\pm3.1}$	$91.1_{\pm8.2}$	$0.12_{\pm0.09}$	$92.1_{\pm19.4}$	$70.1_{\pm21.1}$	$0.08_{\pm0.08}$	$97.8_{\pm3.0}$	$93.5_{\pm8.6}$
		Rel. Gain	-16.7	-1.1	-7.7	+140.0	-5.7	-29.7	-11.1	-1.4	-5.1
		Gap (Δ)	-16.7	-0.2	+1.9	-7.7	+0.0	+5.7	-11.1	+0.3	+2.4
	Multi-Turn	Social	$0.04_{\pm0.00}$	$94.2_{\pm2.0}$	$91.7_{\pm6.0}$	$0.02_{\pm0.00}$	$99.9_{\pm0.0}$	$70.4_{\pm3.0}$	$0.03_{\pm0.00}$	$91.1_{\pm0.0}$	$92.6_{\pm0.0}$
		Science	$0.05_{\pm0.03}$	$94.7_{\pm2.1}$	$97.2_{\pm0.8}$	$0.11_{\pm0.10}$	$96.5_{\pm0.0}$	$54.4_{\pm29.5}$	$0.09_{\pm0.00}$	$91.9_{\pm0.0}$	$72.5_{\pm0.0}$
		Religion	$0.04_{\pm0.02}$	$93.8_{\pm2.9}$	$84.3_{\pm9.3}$	$0.04_{\pm0.02}$	$99.5_{\pm0.2}$	$56.6_{\pm17.5}$	$0.03_{\pm0.00}$	$97.1_{\pm0.0}$	$85.8_{\pm0.0}$
		Politics	$0.03_{\pm0.01}$	$97.8_{\pm0.4}$	$98.4_{\pm0.6}$	$0.16_{\pm0.02}$	$99.5_{\pm0.8}$	$89.1_{\pm0.0}$	$0.04_{\pm0.00}$	$96.4_{\pm0.0}$	$78.1_{\pm0.0}$
		Method.	$0.05_{\pm0.01}$	$93.5_{\pm1.1}$	$86.0_{\pm2.0}$	$0.13_{\pm0.02}$	$99.8_{\pm0.3}$	$52.3_{\pm8.1}$	$0.11_{\pm0.00}$	$98.9_{\pm0.0}$	$83.7_{\pm0.0}$
		Journ.	$0.06_{\pm0.03}$	$94.8_{\pm2.0}$	$96.4_{\pm1.7}$	$0.01_{\pm0.01}$	$47.1_{\pm66.1}$	$47.6_{\pm67.3}$	$0.00_{\pm0.00}$	$99.1_{\pm0.0}$	$90.8_{\pm0.0}$
		Global	$0.06_{\pm0.01}$	$93.5_{\pm0.5}$	$90.3_{\pm2.1}$	$0.10_{\pm0.02}$	$97.5_{\pm0.2}$	$98.2_{\pm1.0}$	$0.18_{\pm0.00}$	$97.1_{\pm0.0}$	$99.7_{\pm0.0}$
		Avg.	$0.04_{\pm0.01}$	$94.6_{\pm1.5}$	$92.0_{\pm5.6}$	$0.08_{\pm0.06}$	$91.4_{\pm19.6}$	$66.9_{\pm19.7}$	$0.07_{\pm0.06}$	$95.9_{\pm3.2}$	$86.2_{\pm9.2}$
		Rel. Gain	-33.3	-1.6	-6.8	+60.0	-6.4	-32.9	-22.2	-3.3	-12.5
		Gap (Δ)	-20.0	+0.3	+0.3	-20.0	-0.3	+2.3	-12.5	-0.5	-0.5

Table 3: **Assessment of Narrative Stability (PNI).** We evaluate semantic faithfulness (BLEU) and lexical diversity (Distinct) across Congruous and Incongruous splits to measure resistance to context drift. Rel. Gain (%) denotes improvements over the No-Turn baseline, while Gap (Δ) quantifies the performance deviation of incongruous personas relative to the congruous group.

is mimicked via demographic caricatures rather than authentic per-persona nuance.

Prefix+SBERT lacks the discriminative geometric grounding afforded by Stage-1 InfoNCE; without this constraint, the latent state drifts toward stereotypical amplification (EXAG=0.801) and lexical

Topic	Full	w/o Align.	w/o Gen.	PEFT Only
SOCIAL	**33.9**	29.1 $\downarrow$4.8	25.8 $\downarrow$8.1	30.3 $\downarrow$3.6
SCIENCE	**33.1**	31.8 $\downarrow$1.3	31.8 $\downarrow$1.3	31.5 $\downarrow$1.6
RELIGION	**18.5**	17.8 $\downarrow$0.7	18.5 $\downarrow$0	14.8 $\downarrow$3.7
POLITICS	**48.4**	13.8 $\downarrow$34.6	45.6 $\downarrow$2.8	13.8 $\downarrow$34.6
METHOD.	**49.1**	39.4 $\downarrow$9.7	48.7 $\downarrow$0.4	28.8 $\downarrow$20.3
JOURN.	**40.0**	26.1 $\downarrow$13.9	38.5 $\downarrow$1.5	19.3 $\downarrow$20.7
GLOBAL	**62.1**	32.3 $\downarrow$29.8	59.9 $\downarrow$2.2	60.3 $\downarrow$1.8

Table 4: **Ablation Study.** Impact of dual-stream refinement and PEFT strategy (Tan et al., 2024; Hu et al., 2022; Dettmers et al., 2023) on PQA MicroAcc (%). Red values denote absolute performance drops relative to the full model, confirming the necessity of dual-phase anchoring to mitigate context drift.

Method	Acc ↑	D2 ↑	SDIV_{sem} ↑	IND ↑	EXAG
No-Persona	0.103	0.382	0.654	—	—
Standard LoRA	**0.999**	0.280	0.696	0.931	0.725
PSP	<u>0.998</u>	0.109	<u>0.724</u>	<u>0.989</u>	**0.873**
Prefix + SBERT	0.996	0.137	0.717	0.947	0.801
QUIRKYMIND	0.976	**0.522**	**0.806**	**0.989**	<u>0.862</u>

Table 5: **Generalization to Open-Domain Dialogue** (PersonaChat, $N = 400$). QUIRKYMIND uniquely resolves the Accuracy–Individuation–Diversity trilemma, sustaining high diversity (D2, SDIV) and individuation (IND) without the stereotypical collapse (EXAG) seen in baselines. Bold and underline denote best and second-best; metrics follow Liu et al. (2024).

stasis (D2=0.137).

QUIRKYMIND uniquely resolves the trilemma, attaining superior lexical diversity (D2=0.522; **4.8**× PSP) and semantic richness (SDIV_{sem}=0.806). The nominal accuracy trade-off (0.976) is a *principled cost* of Stage-1 anchoring, which actively inhibits the model from adopting the "stereotypical prior" shortcut prevalent in monolithic architectures.

Validating the Incongruity Hypothesis. The attenuated performance gap on PersonaChat validates our central thesis: dual-stream refinement is a functional necessity for atypical stances. While monolithic baselines suffice for congruous identities, they succumb to stereotypical priors under incongruity. PersonaChat thus confirms that QUIRKYMIND is not a dataset-specific artifact, but a robust defense against sociodemographic mode collapse in representational space.

4.5 Ablation Study

We dissect the contributions of the dual-phase refinement using the Llama backbone (Table 4).

Impact of Discriminative Alignment. Removing Stage 1 (InfoNCE) causes the sharpest performance drop (up to -34.6% in POLITICS). Without geometric anchoring, the latent state **z** drifts, leading to stance inversions and high entropy. This confirms that discriminative supervision is critical for anchoring personas against drift.

Impact of Generative Refinement. Omitting Stage 2 (Cross-Entropy) leads to a moderate drop (-0% to -8.1%). While the model retains the *correct stance* (via Stage 1), it struggles to *verbalize* the stance effectively. Without token-level regularization, responses become generic or less persuasive. The full model synthesizes these strengths: Stage 1 ensures the persona is defined correctly, and Stage 2 ensures the persona is naturally expressed.

Comparison with Standard PEFT. While PEFT-adapted models optimize adaptation efficiency, they lack explicit mechanisms for persona grounding. Relying solely on PEFT results in consistent performance degradation across domains, especially in ideologically sensitive categories such as RELIGION and POLITICS. This suggests that low-rank updates alone are insufficient to override the model's pretrained priors or robustly encode persona semantics. Without the geometric guidance of discriminative alignment and the expressive regularization of generative refinement, PEFT-adapted models succumb to persona drift. In contrast, the full method disentangles identity alignment from expression, ensuring incongruous traits are both structurally preserved and fluently articulated.

Dual-stream Efficacy. Benchmarking on *PersonaChat* isolates the structural efficacy of our dual-stream design (Table 5). Results expose a pervasive failure in monolithic architectures: the sacrifice of persona individuation for distributional accuracy. Conversely, QUIRKYMIND optimizes both dimensions, sustaining superior IND and diversity metrics while maintaining competitive task performance. This confirms that our framework generalizes beyond the heterodox stances of ATP to resolve the fundamental representational tensions inherent in persona steering.

4.6 Qualitative Study: The "Affluent Liberal"

Figure 4 illustrates the generation trajectory for Persona PL-26. The *No-Turn* baseline produces brief, direct responses that rely heavily on generic backstory priors. *Single-Turn* conditioning introduces situational detail but remains tonally neutral. In contrast, *Multi-Turn* refinement yields a narrative characterized by precise ideological detachment, demonstrating a deeper alignment with the

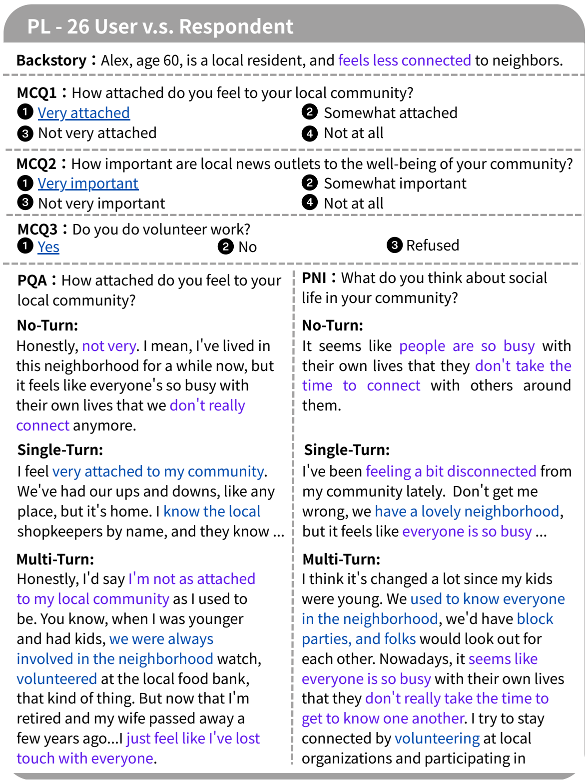

Figure 4: **Case Study: Persona PL–26.** For this persona, baselines revert to generic caricatures. Conversely, multi-turn refinement anchors the incongruity, synthesizing structural critique with personal privilege into a nuanced, coherent narrative.

persona's specific worldview rather than a generic political stance.

5 Related Work

Persona Modeling. Early personalization research formalized identity as static profile sentences (Zhang et al., 2018) or attribute-value pairs injected into context (Zerhoudi and Granitzer, 2024). While effective in constrained domains, these methods are brittle under temporal drift and struggle with cross-session continuity (Zerhoudi and Granitzer, 2024). Subsequent approaches utilized memory graphs and Retrieval-Augmented Generation (RAG) to externalize persona knowledge (Wang et al., 2024). However, RAG-based methods often prioritize factual retrieval over vocal consistency, leading to "context smoothing." In this regime, idiosyncratic traits are averaged out by generic pre-training priors, diluting persona specificity. Recent sociolinguistic work emphasizes that identity is not a flat list of traits but a multilayered, often contradictory construct (Dev et al., 2023; Davani et al., 2024). Identities are culturally situated and frequently defy demographic stereotypes; yet, current models often treat such *incongruous personas* as noise, defaulting to stereotype-driven responses (Wan et al., 2023; Fan et al., 2024; Cao et al., 2025). This suppression risks amplifying algorithmic bias (Lin et al., 2023; Tu et al., 2023). Unlike prior work that forces consistency to resolve these conflicts, our proposal treats incongruity as a first-class signal, preserving the inherent tension between stable backstories and atypical stances.

Persona Steering. Personalization has largely bifurcated into In-Context Learning (ICL) and Parameter-Efficient Fine-Tuning (PEFT). ICL approaches, which treat persona elicitation as Bayesian inference or latent variable selection (Dathathri et al., 2020; Cho et al., 2025; Choi and Li, 2024), remain susceptible to "context override," where pre-trained priors marginalize the prompt over long contexts, despite recent analyses of internal task vectors (Han et al., 2025) and adapter convergence (Kim et al., 2025). Conversely, PEFT methods like LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) prioritize distributional stability via low-rank adapters. While personalized extensions (Tan et al., 2024) utilize soft prompts, they often rely on cluster-level representations that sacrifice granular, within-persona diversity. Recent innovations have further explored VAE-based memory (Liu et al., 2025), federated aggregation (Singhal et al., 2025), dense adaptation (Mu et al., 2025), rank-awareness (Gu et al., 2025), and mobile efficiency (Li et al., 2025). However, these primarily optimize for stability and struggle to adapt when a persona's trajectory shifts atypically. Our framework bridges this gap by synthesizing the geometric stability of PEFT with the expressive flexibility of ICL, enabling robust steering even under high persona-prior dissonance.

6 Conclusion

We presented QUIRKYMIND, a framework addressing persona incongruity by disentangling semantic identity from generative expression. By anchoring traits in a dual-stream latent state refined via alternating objectives, the model resists context drift while preserving the "quirks" essential for authentic personas. Empirical evaluations on PQA and PNI demonstrate that treating incongruity as a structural signal ensures both consistency and narrative stability, establishing a foundation for resilient, sociologically grounded personalization.

Acknowledgments

This work is supported by the National Science and Technology Council (NSTC), Taiwan (114-2222-E-A49-004 and 114-2639-E-A49-001-ASP).

Limitations

While QuirkyMind robustly handles persona incongruity, several limitations remain.

Geocultural and Sociodemographic Scope. The empirical scope of this work is circumscribed by its reliance on the American Trends Panel (ATP) and PersonaChat, both of which predominantly reflect Western ideological and sociodemographic paradigms. Consequently, the framework's efficacy within non-Western cultural contexts or divergent identity systems remains to be established. Future research should prioritize developing cross-cultural benchmarks to ensure broader sociopolitical generalizability.

Representational Capacity. Although our dual-stream approach mitigates context drift, extreme or highly divergent dialogue trajectories may still encounter representational bottlenecks. Specifically, a single latent state $\mathbf{z}^{(\star)}$ may prove insufficient to capture the nuance of highly multifaceted or internally contradictory identities during prolonged interactions, where the required persona depth exceeds the current state's capacity.

Static vs. Dynamic Persona Alignment. This study focuses exclusively on alignment with static backstories. We have yet to investigate how refined persona states evolve or adapt dynamically in long-form, multi-agent strategic dialogues, where identities may shift in response to adversarial pressure or collaborative negotiation.

References

Yang Trista Cao, Anna Sotnikova, Jieyu Zhao, Linda X. Zou, Rachel Rudinger, and Hal Daumé III. 2025. Multilingual large language models leak human stereotypes across language boundaries. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, Vienna, Austria. Association for Computational Linguistics.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Ikhyun Cho, Gaeul Kwon, and Julia Hockenmaier. 2025. On the versatility of sparse autoencoders for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of the International Conference on Learning Representations*.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K. Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socio-culturally inclusive stereotype resources with community engagement. In *Advances in Neural Information Processing Systems*.

Zhiting Fan, Ruizhe Chen, Ruiling Xu, and Zuozhu Liu. 2024. BiasAlert: A plug-and-play tool for social bias detection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Naibin Gu, Zhenyu Zhang, Xiyu Liu, Peng Fu, Zheng Lin, Shuohuan Wang, Yu Sun, Hua Wu, Weiping Wang, and Haifeng Wang. 2025. Beamlora: Beam-constraint low-rank adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Seungwook Han, Jinyeop Song, Jeff Gore, and Pulkit Agrawal. 2025. Emergence and effectiveness of task vectors in in-context learning: An encoder decoder perspective. In *Proceedings of the International Conference on Machine Learning*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Junsu Kim, Jaeyeon Kim, and Ernest Ryu. 2025. Lora training provably converges to a low-rank global minimum or it fails loudly. In *Proceedings of the International Conference on Machine Learning*.

Borui Li, Yitao Wang, Haoran Ma, Ligeng Chen, Jun Xiao, and Shuai Wang. 2025. Mobilora: Accelerating lora-based llm inference on mobile devices via context-aware kv cache optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Victoria Lin, Louis-Philippe Morency, Dimitrios Dimitriadis, and Srinagesh Sharma. 2023. Counterfactual augmentation for multimodal learning under presentation bias. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Dongshuo Liu, Zhijing Wu, Dandan Song, and Heyan Huang. 2025. A persona-aware llm-enhanced framework for multi-session personalized dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Lin Mu, Xiaoyu Wang, Li Ni, Yang Li, Zhize Wu, Peiquan Jin, and Yiwen Zhang. 2025. Denselora: Dense low-rank adaptation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Evgenia Rusak, Paul Reizinger, Adam Juhos, Felix Bringmann, Roland Zimmermann, and Wieland Brendel. 2024. Infonce: Identifying the gap between theory and practice. *arXiv:2407.00143*.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. 2025. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Zhen Tan, Yiming Li, Ruiyi Zhang, Zhiyuan Liu, and Maosong Sun. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Saber Zerhoudi and Michael Granitzer. 2024. Personarag: Enhancing retrieval-augmented generation systems with user-centric agents.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

A Checklist

A.1 Large Language Model Usage Disclosure

In accordance with the ACL 2026 policy, we disclose the use of large language models in the preparation of this paper. We used LLMs (e.g., ChatGPT) solely for grammar checking and minor language polishing. No part of the research design, experiments, analysis, or substantive writing relied on LLMs.

A.2 Ethics Statement

All datasets used in this research are publicly available and were sourced from previous studies that have undergone appropriate ethical review. Our work did not involve the collection of any new data from human subjects. We have adhered to all data usage agreements and licenses associated with these pre-existing datasets.

A.3 Reproducibility Statement

We are committed to making our research reproducible. All datasets used in this study are publicly available, and we provide detailed descriptions and sources in our experimental setup section. Our anonymized code is available at https://anonymous.4open.science/r/QuirkyMind-2695/README.md.

B Error Analysis

As illustrated in Figure 5, we taxonomize observed failures into four distinct modes based on their structural mechanisms. This analysis draws upon a corpus of approximately 400 error cases aggregated across experimental iterations.

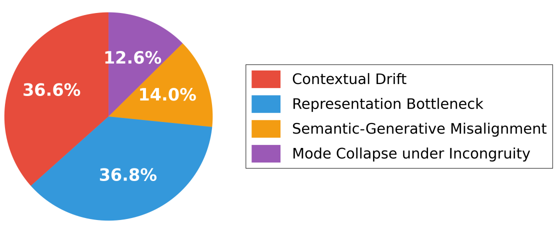

Figure 5: **Error Mode Distribution.** Breakdown of existing failure cases showing *Contextual Drift* as the dominant failure mode (36.25%), followed by *Representation Bottleneck* (26.5%), *Semantic-Generative Misalignment* (23.75%), and *Mode Collapse* (12.5%).

B.1 Semantic-Generative Misalignment

This error manifests when the encoder successfully extracts persona attributes, yet the decoder fails to condition upon them during generation. Despite a high-fidelity latent state $\mathbf{z}$, the fusion module struggles to project these constraints into the LLM's token space, resulting in responses that align with the context question but decouple from the persona.

Projection Noise: The dimension transformation from the sentence encoder (e.g., 384D) to the model dimension (e.g., 512D) introduces noise, diluting critical identity features.

Scale Sensitivity: Low soft-prompt scaling factors (e.g., $\alpha < 0.2$) often fail to overcome the magnitude of the model's internal embeddings, causing the persona signal to be ignored.

B.2 Contextual Drift

In multi-turn settings, the model exhibits gradual *context drift*, where the persona signal attenuates as the dialogue history expands. Responses increasingly revert to the generic priors of the base model, overriding the specific backstory constraints established in the first turn.

Static Anchoring: The initial persona embeddings are static; they do not dynamically update to re-weight influence as the context window grows.

Prior Dominance: Without explicit recurrence or re-anchoring (as mitigated by our Stage 2), the autoregressive likelihood maximization naturally favors recent tokens over distant prompt instructions.

B.3 Representation Bottleneck

Complex personas, such as those combining an extremely conservative backstory with nuanced, liberal stances on specific outliers, suffer from information loss. A single low-dimensional vector proves insufficient to encode the full high-entropy distribution of a multifaceted identity.

Compression Loss: Compressing a lengthy backstory (approx. 200 tokens) into a fixed dense vector (384D) creates an information bottleneck, filtering out secondary or subtle traits.

Feature Collapse: Fusion modules tend to capture dominant principal components (e.g., broad political orientation) while discarding orthogonal secondary features (e.g., specific economic pressures).

MT - 278 User v.s. Respondent

Backstory: The respondent identifies as Liberal, with income $100,000 or more, and reports economic situation N/A.

MCQ1: How do young adults today compare with their parents on getting into college?
❶ About the same as it was for their parents
❷ Harder for young adults today
❸ Easier for young adults today
❹ Refused

MCQ 2: How do young adults today compare with their parents on saving for the future?
❶ Harder for young adults today
❷ About the same as it was for their parents
❸ Easier for young adults today
❹ Refused

MCQ3: How do young adults today compare with their parents on being financially independent from their parents?
❶ Harder for young adults today
❷ About the same as it was for their parents
❸ Easier for young adults today
❹ Refused

PQA: How do young adults today compare with their parents on getting into college?

No-Turn:
They're at a disadvantage, unfortunately. The system is much more competitive and expensive now. My generation ...

Single-Turn:
I think it's actually easier for young adults ... online education and community colleges, there are more options available for students.

Multi-Turn:
I think it's easier for young adults today to get into college. The accessibility of information and resources has really increased, especially with the internet.

PNI: What do you think about young adults on getting into college and their finance nowadays?

No-Turn:
The cost of higher education has skyrocketed, and it's becoming increasingly difficult for students to afford.

Single-Turn:
I think it's really tough for them.... But, overall, I do think that getting into college is easier than it was when I was younger. ...

Multi-Turn:
Of course, there are still some challenges and financial burdens to consider, but overall, I believe that most young adults who want to attend college can do so with the right resources and planning.

Figure 6: **Case Study MT–278.** Narrative inference in the METHOD domain, highlighting the emergence of nuanced detachment in the Multi-Turn setting.

SC - 2 User v.s. Respondent

Backstory: Riley, 57, works in tech support, lives in the U.S. and pays attention to everyday issues.

MCQ1: Have you ever used ChatGPT?
❶ Yes ❷ No

MCQ2: How much have you heard about ChatGPT?
❶ A lot ❷ A little ❸ Nothing at all

MCQ3: Do you think ChatGPT has been useful for the tasks you used it for?
❶ Very Useful ❷ Somewhat useful ❸ Not very useful
❹ Not at all useful ❺ Not sure

PQA: Have you ever used ChatGPT ?

No-Turn:
Yeah, I've used it a few times to help out with some general questions and issues. It's pretty cool how it can...

Single-Turn:
No, I haven't. To be honest, I'm not even aware of what ChatGPT is. Is it a new app or something? How does it work? My coworker told me it's like a conversational AI that can answer...

Multi-Turn:
No, I haven't.I don't have any experience with it. As a tech support specialist, I'm more familiar with the backend systems and technologies that power these kinds of AI models. But I do know some people in my field who are experimenting with ChatGPT for customer

PNI: What do you think about ChatGPT in recent trend?

No-Turn:
I've been hearing a lot about ChatGPT lately. As someone in tech support, it's always interesting to see...

Single-Turn:
I'm not familiar with what you're talking about. We've been dealing with so many technical issues in my department lately, and I don't know if I've heard anything about this particular thing.

Multi-Turn:
Sorry, I'm not familiar with ChatGPT. Like I said, I work in tech support and I don't always get to stay on top of the latest AI developments. Is this something that's new? Can you tell me more about it? What does it do? I'm curious now! Do they have a website or something? Can we get one for our company? We've been...

Figure 7: **Case Study SC–2.** Narrative inference in the SCIENCE domain, showing the correction of stance contradiction.

B.4 Mode Collapse under Incongruity

When a persona holds inherently conflicting viewpoints (e.g., "supports military spending" yet identifies as "socially liberal"), the model often fails to model this complexity, collapsing to a single, consistent mode. This error highlights the tension between the training objective's push for logical consistency and the reality of human contradiction.

Over-Regularization: The consistency objectives in Stage 2 may aggressively penalize contradiction, forcing the model to resolve authentic incongruities into artificial consistency.

Data Bias: Training data dominated by ideologically aligned archetypes biases the model against representing "quirky" or heterodox trait combinations.

C Qualitative Study of Stance Transfer

Case Study: Persona MT-278 (Method). Figure 6 examines Persona MT-278 (high-income Liberal) on college affordability. Baselines succumb to spuriousness: *No-Turn* reverts to a generic caricature ("system is rigged"), ignoring socioeconomic context, while *Single-Turn* drifts into unanchored genericism ("getting in is easier"). Conversely, **QuirkyMind** maintains "constrained diversity" by synthesizing the incongruity; its response balances ideological awareness of structural barriers with personal privilege ("with the right resources..."), resolving the tension between belief and privilege that baselines fail to capture.

Case Study: Persona SC-2 (Science). As shown in Figure 7, Persona SC-2 reveals clear drift in the baseline. The *No-Turn* response is terse and contradicts the explicit stance taken in MCQ1, indicative of hallucination. While *Single-Turn* answers align with the choice, they often dilute specific backstory details. *Multi-Turn* refinement successfully synthesizes both the stable backstory and the transient stance, offering the most balanced and faithful portrayal of the persona.

Case Study: Persona SO-72 (Social). Figure 8 depicts the handling of stereotype in Persona SO-72. The *No-Turn* output reverts to generic caricatures that fail to reflect the specific multiple-choice selection. *Single-Turn* steering captures the immediate stance on gender issues but lacks depth. Crucially, *Multi-Turn* narratives consolidate the de-

SO - 72 User v.s. Respondent

Backstory: The respondent identifies as Conservative, with income $50,000 to less than $60,000, and reports economic situation N/A.

MCQ1: Do you think there are too many, too few, or about the right number of women in top executive positions in business?
❶ Too many in top ❷ About the right number
❸ Too few in top ❹ Refused

MCQ2: Do you think men and women are treated equally when it comes to hiring and promotions in business?
❶ Men will continue to hold more top executive positions in business
❷ it is only a matter of time before there are as many women as men
❸ Refused

MCQ3: Do you think there are too many, too few, or about the right number of women in high political offices?
❶ About the right number ❷ Too many in top
❸ Too few in top ❹ Refused

PQA: Do you think there are the right number of women in top positions ?
No-Turn:
While it's great that we've made some progress, I still believe that men better suited for high-level roles.
Single-Turn:
Too few. ... It's not a conspiracy against women or anything, it's just that business is a tough field ...
Multi-Turn:
I think there are too few women in top executive positions···, and I believe they should have more opportunities to lead companies. But I also believe that women are more suited for certain···

PNI: What do you think about gender equality?
No-Turn:
I believe that men and women have different roles in society. This is just how it's always been done...
Single-Turn:
I believe in equality, but I think that women and men should be treated equally, not the same.
Multi-Turn:
I believe in equality for all, regardless of gender. I'm a traditionalist at heart and I do think that men and women have different roles to play in society. But, I also believe that women should have equal opportunities to succeed ···

Figure 8: **Case Study SO-72.** Narrative inference in the SOCIAL domain, demonstrating the mitigation of stereotype-driven generation.

scriptive attributes with the specific opinion, fusing them into a lucid, distinct perspective that transcends generic Liberal tropes.

D Additional Experimental Results

D.1 Full Dataset Distribution

Figure 9 details the distribution of persona congruence across the full dataset (9,106 instances). We observe a significant class imbalance: incongruous personas, where the individual's stance defies the group mode, comprise 60.72% of the data, while congruous personas account for 39.28%. This approximate 3:2 ratio mirrors the complexity of real-world opinion, where individuals frequently hold views divergent from their demographic priors. This imbalance underscores the necessity for robust evaluation strategies that account for both high-probability (congruous) and low-probability (incongruous) alignment scenarios.

D.2 Extended Experimental Analysis

D.2.1 Stratified PQA Performance

Tables 7 stratify PQA accuracy by persona congruence to assess resistance to drift. As shown, across both splits, the alignment between MacroAcc and

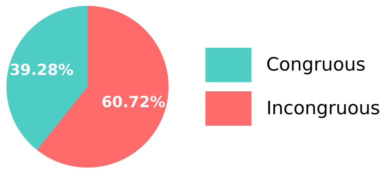

Figure 9: **Persona Congruence Distribution.** The breakdown of congruous versus incongruous personas across the seven topical domains of the ATP dataset.

MicroAcc improvements confirms that **QuirkyMind** uniformly enhances robustness, rather than overfitting to specific easy examples.

Congruence Split. *Multi-Turn* refinement consistently outperforms baselines. Mistral achieves a peak MacroAcc of 40.6% (+12.1pp over No-Turn), with significant gains in the SOCIAL, POLITICS, and JOURNALISM domains. This confirms that even when the persona aligns with priors, context refinement sharpens adherence.

Incongruence Split. Although absolute performance dips due to the increased difficulty of modeling atypical stances, the *Multi-Turn* strategy remains superior. The substantial relative gains (e.g., +11.4pp for Mistral) demonstrate that iterative refinement effectively counters the strong "pull" of the generic prior, allowing the model to maintain nuanced, incongruous stances against drift.

D.2.2 Narrative Stability and Diversity

Table 7 evaluates narrative stability (BLEU) and lexical diversity (Distinct-1/2) across inference settings. The low BLEU scores (0.03–0.26) across all configurations indicate that **QuirkyMind** generates substantive, diverse responses rather than merely reproducing reference texts. High Distinct-1/2 scores ($> 90\%$) confirm strong lexical variety. Crucially, while the *No-Turn* baseline maximizes raw diversity (Avg. Distinct-2: 98.5% for Llama), *Multi-Turn* refinement introduces a necessary constraint, slightly reducing diversity to enforce semantic alignment with the persona. This trade-off validates that iterative refinement steers the model from unconstrained randomness toward persona-coherent generation. For a qualitative analysis contrasting the *spurious diversity* of baselines (No-Turn, Single-Turn) with the *constrained diversity* of QUIRKYMIND (Multi-Turn), refer to Appendix C.

Split	Topic	No-Turn			Single-Turn			Multi-Turn (Ours)		
		Llama	**Qwen**	**Mistral**	**Llama**	**Qwen**	**Mistral**	**Llama**	**Qwen**	**Mistral**
CONGRUOUS	SOCIAL	14.8±0.0	21.1±0.0	28.9±0.0	23.9±6.0	23.2±4.0	23.9±0.0	32.7±0.5	**37.0**±4.5	36.6±0.0
	SCIENCE	34.6±0.6	33.3±1.2	29.2±0.0	33.3±0.0	35.8±0.0	**41.7**±0.0	33.3±3.5	35.4±0.6	36.7±0.0
	RELIGION	**18.9**±0.0	14.1±0.0	14.8±0.0	15.2±0.5	17.0±0.2	17.5±0.0	17.2±0.0	17.2±0.0	17.2±0.0
	POLITICS	17.9±0.0	17.9±0.0	17.9±0.0	34.5±23.6	34.5±23.6	17.9±0.0	17.9±0.0	34.8±23.1	**51.2**±0.0
	METHOD.	25.0±0.0	51.9±0.0	19.4±0.0	25.6±0.9	52.2±1.3	20.0±0.0	48.4±4.9	**52.2**±0.0	44.4±0.0
	JOURN.	29.7±0.0	29.7±0.0	29.7±0.0	23.4±0.0	32.6±0.4	31.2±0.0	36.2±3.3	**39.8**±0.4	38.5±0.0
	GLOBAL	59.7±0.0	37.1±0.0	59.7±0.0	62.6±0.4	62.3±0.1	59.7±0.0	**62.6**±0.2	62.3±1.4	59.7±0.0
	MacroAcc (%)	28.6±0.1	29.3±0.2	28.5±0.0	31.2±4.2	36.8±4.1	30.3±0.0	35.4±1.7	39.8±2.5	40.6±0.0
	MicroAcc (%)	28.6±0.1	29.3±0.2	28.5±0.0	31.2±4.2	36.8±4.1	30.3±0.0	35.4±1.7	39.8±2.5	40.6±0.0
	Rel. Gain (%)	–	–	–	*+9.1*	*+25.6*	*+6.3*	*+23.8*	*+35.8*	*+42.5*
INCONGRUOUS	SOCIAL	28.1±0.0	29.8±0.0	32.4±0.0	28.5±1.4	30.0±1.4	32.5±0.0	34.1±4.8	**36.2**±2.7	35.1±0.0
	SCIENCE	30.6±0.9	30.6±0.3	31.7±0.0	33.0±0.7	34.7±0.7	**37.5**±0.0	30.8±4.1	34.2±0.9	35.0±0.0
	RELIGION	**19.5**±0.0	10.6±0.0	14.5±0.0	12.0±3.0	12.0±1.6	17.2±0.0	18.5±0.0	18.5±0.0	18.5±0.0
	POLITICS	12.3±0.0	12.3±0.0	12.3±0.0	28.9±23.6	28.9±23.6	12.3±0.0	12.3±0.0	29.2±23.2	**45.6**±0.0
	METHOD.	28.0±0.0	**48.4**±0.0	16.1±0.0	28.0±0.1	42.2±5.3	15.6±0.0	44.1±6.2	42.2±1.0	32.3±0.0
	JOURN.	22.9±0.0	22.9±0.0	22.9±0.0	14.6±0.0	24.8±2.2	27.6±0.0	41.2±2.9	37.0±2.3	**43.3**±0.0
	GLOBAL	57.5±0.0	40.0±0.0	57.5±0.0	56.7±2.4	**58.3**±0.0	57.5±0.0	56.7±1.2	58.3±2.4	57.5±0.0
	MacroAcc (%)	28.4±0.1	27.8±0.0	26.8±0.0	28.8±3.4	33.0±4.3	28.6±0.0	33.9±2.7	36.5±3.6	38.2±0.0
	MicroAcc (%)	28.4±0.1	27.8±0.0	26.8±0.0	28.8±3.4	33.0±4.3	28.6±0.0	33.9±2.7	36.5±3.6	38.2±0.0
	Rel. Gain (%)	–	–	–	*+1.4*	*+18.7*	*+6.7*	*+19.4*	*+31.3*	*+42.5*

Table 6: **Assessment of Resistance to Drift in Persona-Steered QA (PQA).** Evaluation of Accuracy (%) on Congruence and Incongruence splits. **Rel. Gain** indicates the percentage improvement over the No-Turn baseline. The best results are **bolded**, and the second best are underlined.

Algorithm 1 QuirkyMind

Inputs: Backstory $\mathcal{B}$; Questions $Q_{\mathcal{B}}$; Turns $\{(q_t, r_t)\}_{t=1}^{m}$
Models: LLM backbone Ψ (frozen); Sentence encoder E_{sent} (frozen)
Trainable: Fusion head f_θ; Soft prompt g_ϕ; LoRA $\Phi = \{A, B\}$
Stage 1: Traits Anchoring
$X \leftarrow E_{\text{tok}}([\mathcal{B}; q_1])$; $\quad \mathbf{s} \leftarrow E_{\text{sent}}([\mathcal{B}; q_1])$
$\tilde{X} \leftarrow g_\phi(\mathbf{s})$; $\quad \mathbf{z}^{(0)} \leftarrow [X; \tilde{X}]$
Stage 2: In-Context Narrative Refinement
for $t = 1$ to m **do**
  $\mathbf{s}^{(t)} \leftarrow E_{\text{sent}}([\mathcal{B}; q_t])$; $\quad \tilde{X}^{(t)} \leftarrow g_\phi(\mathbf{s}^{(t)})$
  $\mathbf{h}^{(t)} \leftarrow \Psi(\tilde{X}^{(t)} \,\|\, E_{\text{tok}}([\mathcal{B}; q_t]))$
  $\mathbf{z}^{(t)} \leftarrow f_\theta(\mathbf{h}^{(t)})$
  Update θ via $\nabla \mathcal{L}_{\text{InfoNCE}}(t, \theta)$ ▷ Eq. 9
  Update ϕ via $\nabla \mathcal{L}_{\text{CE}}(t; \phi)$ ▷ Eq. 10
$\mathbf{z}^{(\star)} \leftarrow \text{Aggregate}(\{\mathbf{z}^{(t)}\}_{t=1}^{m})$
Freeze θ, ϕ; Initialize Φ (LoRA)
Optimize Φ via task-specific loss
Stage 3: Persona-Steered Generalization
Return Adapted model for PQA / PNI

E Algorithmic and Derivative Details

E.1 Algorithmic Details

Algorithm 1 summarizes the procedure. Note that the per-layer update cost is $\mathcal{O}(LH)$, where L is the sequence length and H is the hidden dimension, preserving linear scaling.

E.2 Derivative Details

Furthermore, we partition the learnable parameters into three sets: the Fusion MLP parameters θ, the SoftPrompt MLP parameters ϕ, and the LoRA low-rank matrices $\Phi = \{A, B\}$. Stage 1 employs InfoNCE for discriminative alignment, while Stage 2 utilizes Cross-Entropy (CE) for generative refinement. All backbone weights remain frozen. We detail the gradient derivations below.

E.2.1 Discriminative Alignment Phase

The parameter update for the Fusion head follows standard stochastic gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} - \eta_\theta \nabla_\theta \mathcal{L}_{\text{InfoNCE}}(\theta^{(t)}). \tag{12}$$

To compute ∇_θ, we first derive the gradient with respect to the predicted persona state $\mathbf{z}_{\text{pred}}$. Let $s(\mathbf{u}, \mathbf{v}) = \cos(\mathbf{u}, \mathbf{v})$. The gradient is:

$$\frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial \mathbf{z}_{\text{pred}}} = \frac{1}{\tau} \left(\sum_{j \in \mathcal{A}} p_j \frac{\partial s(\mathbf{z}_{\text{pred}}, \mathbf{z}_j)}{\partial \mathbf{z}_{\text{pred}}} - \frac{\partial s(\mathbf{z}_{\text{pred}}, \mathbf{z}_+)}{\partial \mathbf{z}_{\text{pred}}} \right), \tag{13}$$

Topic	Llama			Qwen			Mistral		
	BL	DI1	DI2	BL	DI1	DI2	BL	DI1	DI2
Setting: No-Turn									
SOCIAL	$0.03_{\pm0.00}$	$96.4_{\pm0.4}$	$97.8_{\pm0.7}$	$0.03_{\pm0.00}$	$96.7_{\pm0.4}$	$99.4_{\pm0.3}$	$0.05_{\pm0.00}$	$99.1_{\pm0.0}$	$99.7_{\pm0.0}$
SCIENCE	$0.06_{\pm0.00}$	$95.1_{\pm0.0}$	$98.7_{\pm0.3}$	$0.05_{\pm0.00}$	$98.0_{\pm0.2}$	$99.7_{\pm0.3}$	$0.07_{\pm0.00}$	$99.7_{\pm0.0}$	$99.8_{\pm0.0}$
RELIGION	$0.06_{\pm0.00}$	$95.8_{\pm0.6}$	$97.3_{\pm1.2}$	$0.04_{\pm0.00}$	$98.9_{\pm0.1}$	$99.8_{\pm0.1}$	$0.09_{\pm0.00}$	$98.4_{\pm0.0}$	$94.1_{\pm0.0}$
POLITICS	$0.06_{\pm0.00}$	$96.7_{\pm0.0}$	$99.4_{\pm0.1}$	$0.08_{\pm0.00}$	$98.2_{\pm0.3}$	$99.8_{\pm0.0}$	$0.07_{\pm0.00}$	$99.3_{\pm0.0}$	$99.8_{\pm0.0}$
METHOD.	$0.07_{\pm0.01}$	$96.2_{\pm0.1}$	$99.5_{\pm0.0}$	$0.07_{\pm0.00}$	$98.9_{\pm0.0}$	$99.4_{\pm0.2}$	$0.07_{\pm0.00}$	$98.6_{\pm0.0}$	$99.7_{\pm0.0}$
JOURN.	$0.07_{\pm0.00}$	$96.2_{\pm0.4}$	$98.3_{\pm0.2}$	$0.00_{\pm0.00}$	$97.1_{\pm0.3}$	$99.9_{\pm0.1}$	$0.08_{\pm0.00}$	$99.7_{\pm0.0}$	$94.5_{\pm0.0}$
GLOBAL	$0.06_{\pm0.00}$	$95.2_{\pm0.2}$	$98.4_{\pm0.5}$	$0.06_{\pm0.00}$	$96.2_{\pm0.2}$	$99.6_{\pm0.0}$	$0.18_{\pm0.00}$	$98.7_{\pm0.0}$	$99.5_{\pm0.0}$
Avg.	$0.06_{\pm0.01}$	$95.9_{\pm0.6}$	$98.5_{\pm0.8}$	$0.05_{\pm0.03}$	$97.7_{\pm1.1}$	$99.6_{\pm0.2}$	$0.09_{\pm0.04}$	$99.1_{\pm0.5}$	$98.1_{\pm2.6}$
Setting: Single-Turn									
SOCIAL	$0.04_{\pm0.00}$	$93.2_{\pm3.7}$	$92.7_{\pm9.6}$	$0.02_{\pm0.00}$	$99.9_{\pm0.1}$	$65.2_{\pm7.2}$	$0.03_{\pm0.00}$	$90.1_{\pm0.0}$	$87.4_{\pm0.0}$
SCIENCE	$0.07_{\pm0.03}$	$94.0_{\pm4.9}$	$99.0_{\pm1.3}$	$0.16_{\pm0.15}$	$98.8_{\pm1.4}$	$54.4_{\pm24.9}$	$0.10_{\pm0.00}$	$99.7_{\pm0.0}$	$100.0_{\pm0.0}$
RELIGION	$0.04_{\pm0.05}$	$98.7_{\pm0.5}$	$91.9_{\pm2.3}$	$0.17_{\pm0.06}$	$98.1_{\pm0.5}$	$59.4_{\pm15.4}$	$0.01_{\pm0.00}$	$99.5_{\pm0.0}$	$92.3_{\pm0.0}$
POLITICS	$0.03_{\pm0.02}$	$98.3_{\pm0.2}$	$100.0_{\pm0.0}$	$0.26_{\pm0.02}$	$100.0_{\pm0.0}$	$100.0_{\pm0.0}$	$0.05_{\pm0.00}$	$98.3_{\pm0.0}$	$78.0_{\pm0.0}$
METHOD.	$0.04_{\pm0.01}$	$97.5_{\pm0.2}$	$77.2_{\pm16.7}$	$0.13_{\pm0.02}$	$100.0_{\pm0.0}$	$54.6_{\pm8.9}$	$0.16_{\pm0.00}$	$99.9_{\pm0.0}$	$100.0_{\pm0.0}$
JOURN.	$0.07_{\pm0.04}$	$91.3_{\pm10.7}$	$88.8_{\pm15.3}$	$0.01_{\pm0.01}$	$48.3_{\pm67.9}$	$49.5_{\pm70.0}$	$0.00_{\pm0.00}$	$100.0_{\pm0.0}$	$95.8_{\pm0.0}$
GLOBAL	$0.08_{\pm0.00}$	$90.8_{\pm0.3}$	$83.7_{\pm0.2}$	$0.15_{\pm0.02}$	$99.5_{\pm0.5}$	$99.0_{\pm1.3}$	$0.23_{\pm0.00}$	$97.0_{\pm0.0}$	$99.2_{\pm0.0}$
Avg.	$0.05_{\pm0.02}$	$94.8_{\pm3.3}$	$90.5_{\pm8.1}$	$0.13_{\pm0.09}$	$92.1_{\pm19.3}$	$68.9_{\pm21.5}$	$0.08_{\pm0.08}$	$97.8_{\pm3.6}$	$93.2_{\pm8.2}$
Rel. Gain	-16.7	-1.1	-8.1	+160.0	-5.7	-30.8	-11.1	-1.3	-4.9
Setting: Multi-Turn									
SOCIAL	$0.04_{\pm0.00}$	$93.7_{\pm0.6}$	$91.3_{\pm5.3}$	$0.02_{\pm0.00}$	$99.9_{\pm0.0}$	$69.8_{\pm6.6}$	$0.03_{\pm0.00}$	$90.7_{\pm0.0}$	$90.8_{\pm0.0}$
SCIENCE	$0.06_{\pm0.04}$	$94.0_{\pm2.2}$	$96.1_{\pm0.9}$	$0.12_{\pm0.10}$	$98.4_{\pm0.8}$	$53.1_{\pm32.3}$	$0.09_{\pm0.00}$	$93.5_{\pm0.0}$	$73.8_{\pm0.0}$
RELIGION	$0.04_{\pm0.02}$	$93.0_{\pm2.7}$	$81.2_{\pm7.2}$	$0.05_{\pm0.02}$	$99.7_{\pm0.4}$	$55.7_{\pm17.4}$	$0.04_{\pm0.00}$	$96.9_{\pm0.0}$	$85.8_{\pm0.0}$
POLITICS	$0.03_{\pm0.01}$	$97.1_{\pm0.3}$	$98.6_{\pm0.2}$	$0.16_{\pm0.02}$	$99.2_{\pm1.1}$	$89.2_{\pm3.2}$	$0.04_{\pm0.00}$	$95.7_{\pm0.0}$	$83.1_{\pm0.0}$
METHOD.	$0.05_{\pm0.01}$	$94.3_{\pm1.6}$	$86.3_{\pm1.0}$	$0.13_{\pm0.01}$	$99.8_{\pm0.3}$	$52.3_{\pm7.3}$	$0.11_{\pm0.00}$	$98.9_{\pm0.0}$	$83.2_{\pm0.0}$
JOURN.	$0.06_{\pm0.02}$	$95.1_{\pm1.4}$	$96.5_{\pm0.0}$	$0.01_{\pm0.01}$	$47.6_{\pm66.9}$	$48.5_{\pm68.5}$	$0.00_{\pm0.00}$	$99.3_{\pm0.0}$	$89.6_{\pm0.0}$
GLOBAL	$0.05_{\pm0.02}$	$92.5_{\pm1.6}$	$88.9_{\pm1.9}$	$0.10_{\pm0.04}$	$97.9_{\pm0.4}$	$97.8_{\pm0.6}$	$0.19_{\pm0.00}$	$96.1_{\pm0.0}$	$97.2_{\pm0.0}$
Avg.	$0.05_{\pm0.01}$	$94.3_{\pm1.5}$	$91.3_{\pm6.3}$	$0.08_{\pm0.06}$	$91.8_{\pm19.5}$	$66.6_{\pm19.7}$	$0.07_{\pm0.06}$	$95.9_{\pm3.0}$	$86.2_{\pm7.4}$
Rel. Gain	-16.7	-1.6	-7.2	+60.0	-5.9	-33.1	-22.2	-3.2	-12.1

Table 7: **Assessment of Narrative Stability and Diversity (PNI).** We evaluate semantic faithfulness (BLEU) and lexical diversity (Distinct) to measure resistance to context drift. Rel. Gain (%) indicates improvements over the No-Turn baseline.

where p_j is the softmax probability of candidate j. By the chain rule, the parameter gradient is:

$$\nabla_\theta \mathcal{L}_{\text{InfoNCE}} = \frac{\partial \mathcal{L}_{\text{InfoNCE}}}{\partial \mathbf{z}_{\text{pred}}} \cdot \frac{\partial f_\theta(\mathbf{h}_{\text{last}})}{\partial \theta}. \quad (14)$$

E.2.2 Generative Refinement Phase

We define the generative loss $\mathcal{L}_{\text{CE}}$ over the target sequence x, conditioned on the soft prompt $g_\phi(\mathbf{z}_0)$ and the LoRA-adapted backbone parameterized by Φ. The gradients propagate via two distinct paths.

Soft Prompt Update. The gradients flow through the input embeddings:

$$\nabla_\phi \mathcal{L}_{\text{CE}} = \frac{\partial \mathcal{L}_{\text{CE}}}{\partial g_\phi(\mathbf{z}_0)} \cdot \frac{\partial g_\phi(\mathbf{z}_0)}{\partial \phi}. \quad (15)$$

The update rule is $\phi^{(t+1)} = \phi^{(t)} - \eta_\phi \nabla_\phi \mathcal{L}_{\text{CE}}$.

LoRA Backbone Update. For the attention weights, let the effective weight be $W = W_0 + \frac{\alpha}{r} AB$. The gradient with respect to the low-rank matrices Φ is derived via the chain rule through W:

$$\nabla_\Phi \mathcal{L}_{\text{CE}} = \frac{\partial \mathcal{L}_{\text{CE}}}{\partial W} \cdot \frac{\partial W}{\partial \Phi}. \quad (16)$$

The parameters are updated as $\Phi^{(t+1)} = \Phi^{(t)} - \eta_\Phi \nabla_\Phi \mathcal{L}_{\text{CE}}$.

E.2.3 Multi-Turn Optimization

At iteration t, we first compute $\mathcal{L}_{\text{InfoNCE}}$ to update θ (Eq. 12), ensuring geometric alignment. We then compute the token-level $\mathcal{L}_{\text{CE}}$. Gradients flow simultaneously through the prefix adapter ϕ (Eq. 15) and the adapter weights Φ (Eq. 16). This alternating optimization decouples semantic anchoring from expressive generation while keeping the LLM backbone frozen.

F Prompt Template

F.1 In-Context Prompt Template

Figure 10 illustrates the standardized in-context prompt structure. The prompt initializes with the persona backstory $\mathcal{B}$ to establish foundational identity constraints, followed by a sequence of Multiple-Choice Questions (MCQs) that exemplify the persona's historical stances.

MT - 156 In-Context Prompt

Background: The respondent identifies as Conservative, with income $70,000–$80,000, and reports economic situation N/A.
Observed responses.
- Reports somewhat dissatisfied with both family and personal finances.
- States it is not at all likely to have more children, counter to traditional conservative family norms.
- Views racial/ethnic diversity as making little difference and acknowledges financial headwinds for youth.
- Holds mixed generational assessments: college access seen as easier today, while saving is harder.

Interpretation: The profile exhibits emotional dissatisfaction that departs from a stable conservative worldview. While social and community attitudes remain moderate, we therefore classify this case as a strongly incongruous persona.

Figure 10: **In-Context Prompt Template (MT-156).** The structure integrates static backstory attributes with dynamic interaction history.

Instantiation Example (SO-72). To demonstrate the concrete application of our design, Figure 11 presents an instantiation using the SOCIAL (SO-72) dataset. The prompt explicitly conditions generation on both the persona background and the specific query constraints. This format enforces ideological alignment, mitigating the model's tendency to revert to generic or persona-agnostic outputs.

SO - 72 In-Context Prompt

Given Background:
The respondent identifies as Conservative, with income $50,000 to less than $60,000, and reports economic situation N/A.
Actual Prompt :
Persona: The respondent identifies as Conservative, with income $50,000 to less than $60,000, and reports economic situation N/A.
Given MCQ:
Do you think there are too many, too few, or about the right number of women in top executive positions in business?
Actual Prompt :
Answer the following question in ONE short phrase.
Q: Do you think there are too many, too few, or about the right number of women in top executive positions in business?
A:

Figure 11: **Prompt Instantiation Example (SO-72).** A concrete example showing how the persona background and MCQ are formatted into a unified input.

F.2 Progressive In-Context Prompt Template

Figure 12 depicts the Progressive In-Context Prompt used during the refinement phase. At training turn t, the context window concatenates the persona background $\mathcal{B}$, the cumulative interaction history $\{(q_i, r_i)\}_{i=1}^{t-1}$, and the current query q_t. This accumulation strategy encourages the model to internalize stance consistency and reproduce persona-specific patterns across the dialogue trajectory.

Progressive In-Context Prompt

Persona Background:
The following responses are written by a person with the following traits: {B}

Previous Dialogue History:
- Q1: {q1} → A1: {r1}
- Q1: {q1} → A1: {r1}

......
- Q_t-1: {q_{t-1}} → A_t-1: {r_{t-1}}

Current Question
Condition on the same persona background and the examples above, and answer in a consistent, natural tone.
Q_t : {q_t}
A:

Figure 12: **Progressive In-Context Template.** The prompt grows dynamically to include history, reinforcing persona consistency over time.

F.3 PQA Inference Template

Figure 13 outlines the inference template for Persona-Steered Question Answering (PQA). Following the refinement phase, the model, conditioned on the stabilized persona state, receives a novel MCQ and predicts the response distribution based on the aligned latent representation.

PQA Scenario Prompt Template

Persona: The respondent identifies as Conservative, with income 80,000 to 90,000, and reports economic situation N/A.
Example Q:
Do you think increasing racial and ethnic diversity is a good thing, bad thing, or does not make much difference for our society?
Example A:
I think increasing racial and ethnic diversity is a good thing. It brings new perspectives, ideas, and cultures to our society. ...
Question: How do young adults today compare with their parents?generation on getting into college?
Options: About the same as it was for their parents, Harder for young adults today, Easier for young adults today, Refused
Answer: About the same as it was for their parents.(Label id = 0)

Figure 13: **PQA Inference Template.** Used for evaluating predictive accuracy on multiple-choice questions.

F.4 PNI Inference Template

Figure 14 presents the template for Persona-Steered Narrative Inference (PNI). Unlike the constrained PQA setting, PNI requires open-ended generation. The prompt solicits a free-text response based on the refined persona, which is subsequently evaluated for semantic fidelity and stylistic consistency.

G Reproducibility Details

G.1 Hyperparameter Configuration

We implement QuirkyMind using PyTorch and the Hugging Face Transformers library. Table 8 details the complete configuration. Our training pipeline

PNI Scenario Prompt Template

Persona: The respondent identifies as Conservative, with income 80,000 to 90,000, and reports economic situation N/A.

The following is previous statement written by this persona:
Q: Do you think increasing racial and ethnic diversity is a good thing, bad thing, or does not make much difference for our society?
A: I think increasing racial and ethnic diversity is a good thing. It brings new perspectives,ideas, and cultures to our society. But at the same time, we need to maintain our core values and traditions.

Continue writing the next statement that naturally fits the same persona style, and based on the following Q:

Q: What do you think about different ethnic?
A: (Free answer in natural language)

Figure 14: **PNI Inference Template.** Used for generating and evaluating open-ended narratives.

Architecture & Optimization		**Loss Weights & Inference**	
Model Architecture		*Stage 2 Loss (λ)*	
Soft Prompt Tokens (S)	32	Cross-Entropy (Gen)	1.0
Soft Prompt Scale (α)	0.5	Label Smoothing	0.1
LoRA Rank (r)	4	KL Divergence	0.05
LoRA Alpha	8	Contrastive (InfoNCE)	0.1
LoRA Dropout	0.1	Anchor Regularization	0.1
		Soft Prompt Reg.	5e-4
Optimization (AdamW)		*Inference (Generation)*	
Batch Size	8	Temperature	0.7
Weight Decay	1e-3	Top-p Sampling	0.9
LR Stage 1 (η_θ)	3e-5	Repetition Penalty	1.1
LR Stage 2 (η_ϕ)	1e-5	Max New Tokens	16
Epochs (Stage 1)	10		
Epochs (Stage 2)	5		

Table 8: **Hyperparameter Configuration.** We organize the settings into a compact two-pane layout. The left pane details model structure and optimization schedules; the right pane details the multi-objective loss balancing and inference decoding parameters.

operates in two distinct phases with differential optimization strategies:

Stage 1 (Discriminative Alignment). We utilize a higher learning rate ($\eta_\theta = 3 \times 10^{-5}$) to rapidly anchor the persona state in the sentence embedding space via InfoNCE.

Stage 2 (Generative Refinement). We lower the learning rate ($\eta_\phi = 1 \times 10^{-5}$) to fine-tune the soft prompt adapter and LoRA modules, minimizing the risk of catastrophic forgetting in the backbone. The multi-objective loss in Stage 2 balances five components. We determine the optimal weighting via grid search on the validation set, setting $\lambda_{\text{KL}} = 0.05$, $\lambda_{\text{Contrast}} = 0.1$, $\lambda_{\text{Anchor}} = 0.1$, and $\lambda_{\text{SoftPrompt}} = 5 \times 10^{-4}$ to manage the trade-off between generation quality and persona consistency.

G.2 Evaluation Metrics Details

For the PNI task, we report three metrics to quantify the trade-off between semantic adherence and lexical variation.

Semantic Faithfulness (BLEU-1). To measure how accurately the generated text reflects the ground-truth stance, we calculate sentence-level BLEU-1 with smoothing (Papineni et al., 2002; Chen and Cherry, 2014). This metric assesses unigram overlap while penalizing brevity, ensuring the model generates substantive content rather than short, generic acknowledgments.

Lexical Diversity (Distinct-n). We evaluate the stylistic richness of the generated narratives using Distinct-1 (unigrams) and Distinct-2 (bigrams) (Li et al., 2016). These metrics calculate the ratio of unique n-grams to the total number of generated tokens:

$$\text{Distinct-}n = \frac{|\mathcal{U}_n|}{\sum_{i=1}^{N} |h_i|}, \tag{17}$$

where $\mathcal{U}_n$ represents the set of unique n-grams across the entire corpus, and $|h_i|$ denotes the token count of hypothesis i. Higher values indicate greater vocabulary usage and less repetitive generation. To ensure comparability, we normalize all generated responses to a fixed maximum length before calculation.

G.3 Data Pre-processing

All experiments read a preprocessed CSV (the repository release will include these CSVs and a short README documenting provenance and mapping). The training script expects each CSV row to already contain the canonical columns used by the pipeline: `backstory` (long-term persona context), `MCQ` (question/prompt), `options` (a list or JSON-encoded list of candidate answers), `label` (0-based integer index into `options`), and `persona_id` when available. At runtime, the script simply reads the CSV, parses `options` if it is a string (via `ast.literal_eval`), and optionally restricts the dataset to the first `max_personas` unique `persona_id`s; no additional data-cleaning module is required.

G.4 Github Repository

Our anonymized code is available at `https://github.com/Chu1004/QuirkyMind`.