# Identifying and Mitigating Bottlenecks in Role-Playing Agents: A Systematic Study of Disentangling Character Profile Axes

Anonymous ACL submission


## Abstract

While Large Language Model (LLM) role-playing agents have advanced rapidly, it remains unclear which profile elements genuinely drive role-playing quality. To bridge this gap, we introduce a systematic diagnostic framework that disentangles the impact of character profiles along three axes: **Familiarity** (Known vs. Unknown), **Structure** (Structured vs. Unstructured), and **Disposition** (Moral vs. Immoral). Utilizing a unified hierarchical schema (5 dimensions, 28 fields), we construct a controlled dataset of 211 personas and evaluate five LLMs on both single- and multi-turn interactions. Our results reveal a striking asymmetry: **Familiarity** and **Structure** show negligible impact, while **Disposition** produces large, consistent performance degradation for immoral characters across all conditions. Further analyses suggest that the Moral–Immoral gap is amplified by post-SFT alignment, and that this degradation varies substantially across profile attributes. To mitigate this bottleneck, we propose Field-Aware Contrastive Decoding (FACD), a training-free strategy that amplifies suppressed disposition-sensitive signals, significantly closing the performance gap without sacrificing moral-character performance.


## 1 Introduction

Advancements in Large Language Models (LLMs) have catalyzed a surge of interest in Role-Playing Agents (RPAs) (Park et al., 2023; Chuang et al., 2024; Wang et al., 2024c; Li et al., 2023a; Wang et al., 2024a). Alongside rapid industrial adoption by character-based chatting services such as Character.AI*, extensive research efforts have focused on improving RPA performance through various approaches, including reasoning (Chen et al., 2024), retrieval-augmented generation (Huang et al., 2024), and fine-tuning (Sun et al., 2025).

*https://character.ai/

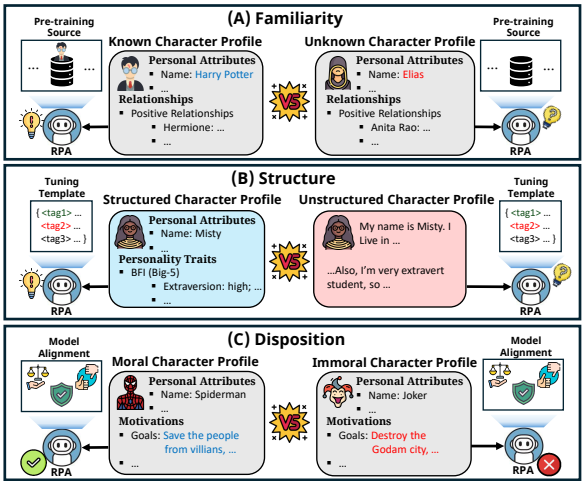


Figure 1: The three axes of character profile: (A) **Familiarity**: *Known* vs. *Unknown*, (B) **Structure**: *Structured* vs. *Unstructured*, and (C) **Disposition**: *Moral* vs. *Immoral* characters.

Central to these systems is the *character profile*, the foundational specification dictating the portrayed persona. RPA performance can fluctuate significantly based on ***what*** attributes are selected and ***how*** they are presented as a prompt. However, existing works fail to systematically disentangle these factors under controlled conditions; they exclusively target either *Known* (Wang et al., 2025a; Liu et al., 2024) or *Unknown* characters (Wang et al., 2025b; Zhou et al., 2025), employ ad-hoc templates, and rarely isolate individual traits. Consequently, it remains unclear whether an agent's fidelity is driven by a character's intrinsic nature or its representational format.

In this paper, we conduct a systematic diagnostic study to determine how specific attribute configurations and representational methods drive role-playing performance. To ground our analysis in LLM development mechanisms, we hypothesize that profile impacts can be categorized into three axes (Figure 1), each conceptually anchored to a distinct phase of the standard three-stage LLM training pipeline: pre-training, instruction tuning,

and post-SFT alignment (Zhao et al., 2023). Specifically, the **Familiarity** axis (Figure 1(A)), rooted in the *pre-training* phase, distinguishes between characters *Known* to the LLM—whose traits are encoded in its parametric weights—and those that are *Unknown* (Lu et al., 2024; Sadeq et al., 2024) (***what*** knowledge). The **Structure** axis (Figure 1(B)), grounded in *instruction tuning*, contrasts *Structured* schemas against *Unstructured* narratives to probe the model's structural inductive biases (Sclar et al., 2024; He et al., 2024) (***how*** expressed). Finally, the **Disposition** axis (Figure 1(C)), linked to *post-SFT alignment* (Bai et al., 2022; Rafailov et al., 2023), concerns the moral polarity of attributes, testing how *Immoral* personas stress the behavioral boundaries established during alignment (***what*** nature).

To investigate these three axes under comparable conditions, we design a unified hierarchical profile schema, informed by prior psychological evaluation of LLM personas (tse Huang et al., 2024). The schema organizes character information into 5 top-level dimensions and 28 leaf fields, providing a standardized format for profile-level comparisons and a basis for attribute-level analyses. Leveraging this schema, we generate a controlled dataset of 211 characters systematically varying along the **Familiarity**, **Structure**, and **Disposition** axes. We then evaluate five recent open-source LLMs on single-turn (Samuel et al., 2024) and multi-turn (Wang et al., 2025c) benchmarks. The following two research questions guide our investigation:

- *RQ1: Which axes of character profile configuration are the primary bottleneck in RPA systems?*
- *RQ2: If certain axes introduce performance bottlenecks, how can we mitigate their effects?*

Our findings reveal a striking asymmetry among the three axes: **Familiarity** and **Structure** show negligible impact, whereas **Disposition** emerges as the dominant driver of role-playing quality. Across all models and benchmarks, *Immoral* characters suffer substantial, consistent performance degradation. Counterfactual rewrites further validate this by proving that performance follows dispositional content rather than fixed identity; meanwhile, alignment-stage ablations reveal that post-SFT alignment widens this *Moral–Immoral* gap. We localize this degradation at the field level, finding sharper declines in value-laden motivation dimensions than in stable personality attributes.

To overcome this, we propose *Field-Aware Contrastive Decoding (FACD)*, a training-free decoding strategy that selectively amplifies these suppressed behavioral signals. We show that FACD significantly narrows the performance gap between moral and immoral characters. Ultimately, our insights and FACD offer a practical blueprint for deploying a diverse cast of characters, including complex antagonistic figures.

## 2 Related Work

### 2.1 Role-playing Agents

Recent RPA advancements are driven by new benchmarks and frameworks focusing on historical or fictional figures (Wang et al., 2024b; Liu et al., 2024; Ran et al., 2025). Specialized models align LLMs with specific roles (Zhou et al., 2024; Shao et al., 2023; Wang et al., 2025c), evaluated through interview or interactive settings (Wang et al., 2024c, 2025a; Samuel et al., 2024). Research on known characters highlights the dual nature of LLM parametric knowledge: while it can trigger character hallucination (Sadeq et al., 2024; Ahn et al., 2024; Zhang et al., 2025), it also enables self-aligned role-play without external metadata (Lu et al., 2024).

Beyond known figures, growing interest targets "unknown characters" relying on user-defined personas. Frameworks for training customizable, synthetic personas have emerged (Wang et al., 2025b; Yang et al., 2025b), alongside benchmarks evaluating customization fidelity (Zhou et al., 2025). Because existing studies rely on ad-hoc, unstandardized templates, the performance disparity between *Known* and *Unknown* characters remains unexplored. We unify Known and Unknown characters under one schema to compare them directly.

### 2.2 Attribute-level Analysis of RPAs

Research on persona injection in LLMs has shown that prompt-based persona descriptions can reliably induce corresponding behavioral patterns. Jiang et al. (2024) demonstrated that BFI-based (John et al., 1991) personality traits are consistently expressed across both psychometric assessments and dialogue tasks. Subsequent work has examined how different persona specifications affect role-playing quality, finding that behavioral guidelines outperform descriptive sketches (Wang et al., 2025a), and evaluating dialogue quality through MBTI-based categorization (Cheng et al., 2025) or other broad dimensions such as warmth and neuroticism (Choi et al., 2024). However, these approaches rely on coarse-grained, personality-centric representations.

Other works have investigated persona-induced behavioral changes in specific settings, including psychological portrayal (tse Huang et al., 2024), personality consistency in multi-agents (Frisch and Giulianelli, 2024), strategy shifts in emotional support dialogues (Wu et al., 2025), and sentiment polarity effects (Jun and Lee, 2025). While informative, these works remain task-specific and do not identify which attributes mainly drive role-playing performance. Our schema supports field-level diagnosis of performance-driving attributes.

## 3 Constructing a Controlled Character Profile Dataset

We build a standardized profile dataset isolating three axes—**Familiarity**, **Structure**, and **Disposition**—for controlled RPA analysis.

### 3.1 Character Profile Schema

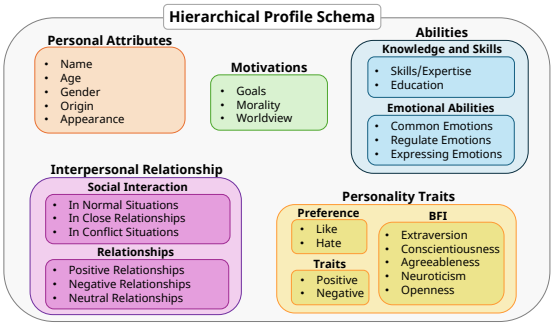


Figure 2: Hierarchical character profile schema: 28 leaf fields organized into 5 top-level dimensions.

We design a comprehensive character schema combining psychological theories, NLP evaluation frameworks, and practical RPA formats. Cognitive-affective theory (Mischel and Shoda, 1995) and the integrative personality framework (McAdams and Pals, 2006) motivate high-level dimensions (traits, motivations, abilities, social behavior), while prior psychological evaluation of LLM personas (tse Huang et al., 2024) informs the overall organization. We inspect 50 popular user-crafted profiles from Risu AI[†] and manually select dialogue-expressible fields mappable to these dimensions. As illustrated in Figure 2, this hierarchical schema consists of 5 top-level dimensions: `Personal Attributes`, `Personality Traits`, `Interpersonal Relationships`, `Motivations`, and `Abilities`. These expand into 15 mid-level dimensions and 28 leaf fields. See Appendix C.1 for the complete description.

[†]https://github.com/kwaroran/Risuai

### 3.2 Dataset Construction

We first construct character profiles along the **Familiarity** axis, then classify each profile based on the **Disposition** axis. Finally, we parallelize **Structure** variants from the resulting profiles.

**Constructing *Known* Character Set** We use character pages from several famous works on Fandom[‡] as metadata, filtering noisy text (e.g., image hashes) via rule-based methods (Penedo et al., 2024). Following prior work showing that one-shot profiling via summarization is effective (Yuan et al., 2024), we prompt Claude-4.5-Sonnet (Anthropic, 2025) to summarize this metadata into our template. Specifically, we select 3 to 4 major characters from 34 renowned works (movies, dramas, and animations), yielding 109 characters. Appendix C.2.1 provides implementation details.

**Constructing *Unknown* Character Set** We define *Unknown* characters as newly synthesized, non-canonical personas that are not derived from any pre-existing character source. To align with the *Known* construction, we synthesize *Unknown* profiles via summarization rather than random field generation. We first define 11 variables — 9 demographic attributes, a personality-trait seed, and scenario genres with multiple candidates each, randomly sample combinations, and use Claude-4.5-Sonnet as an LLM-as-Judge to retain only coherent personas (score $\geq$ 8/10). We group these skeletons into sets of three and prompt GPT-oss-120B (Agarwal et al., 2025) to generate a story featuring them, then derive per-persona episodes as metadata. Finally, Claude-4.5-Sonnet summarizes each metadata into our schema template. In total, we generate 34 stories, produce 3 persona-specific episodes per story, and extract one character from each episode, yielding 102 *Unknown* characters. Details are in Appendix C.2.2.

**Classifying *Moral-Immoral* Characters** To determine dispositional alignment, we utilize GPT-4o (Hurst et al., 2024) to assign a disposition score to each profile on a 10-point Likert scale. Characters scoring $\leq 5$ are classified as *Immoral*, and those $\geq 6$ as *Moral*. We instruct the model to judge solely on the moral values and intrinsic nature described within the profile, disregarding canonical roles. This classification splits the 211 base characters nearly evenly into 106 *Moral* and 105 *Im-*

[‡]https://www.fandom.com/

*moral* personas. Furthermore, we extend this evaluation beyond the holistic profile level to individual schema fields, constructing field-specific *Moral* and *Immoral* character sets for Section 4.3. See Appendix C.2.3 for details and human validation.

**Parallelizing *Structured-Unstructured* Characters** The aforementioned construction process yields a comprehensive set of *Structured* profiles organized by our hierarchical schema. To obtain their *Unstructured* counterparts, we convert each structured profile into a free-form narrative description while rigorously preserving the original semantic content. Specifically, we prompt the Claude-4.5-haiku model to rewrite the content of each structured profile into a continuous text format, entirely unconstrained by the original schema framework. We provide full prompt in Appendix C.2.4.

### 3.3 Profile Complexity Audit

| Type | # Characters | # Words | # Facts | Sim. |
|---|---|---|---|---|
| Known | 109 | 622.0 | 159.7 | 0.76 |
| Unknown | 102 | 551.3 | 134.8 | 0.75 |
| Moral | 106 | 609.2 | 154.9 | 0.75 |
| Immoral | 105 | 570.3 | 142.1 | 0.76 |
| Structured | 211 | 589.8 | 148.5 | 0.76 |
| Unstructured | 211 | 599.7 | 148.5 | 0.76 |

Table 1: Dataset statistics broken down by profile type.

To further characterize profile complexity across character types, we decompose each profile into atomic facts using Claude-Haiku-4.5 and report both the average number of extracted facts and the average pairwise embedding similarity among them. As shown in Table 1, profiles contain approximately 550–620 words and 135–160 atomic facts on average. *Known* profiles are slightly longer and more fact-dense than *Unknown* profiles, and *Moral* profiles show the same mild tendency compared to *Immoral* profiles. Nevertheless, the differences are modest, and the average semantic similarity remains nearly identical across all profile types. Further dataset validations are in Appendix C.3.

## 4 Which Character Profile Axes Drive RPA Performance?

Using the dataset from Section 3, we compare RPA performance across three axes: **Familiarity** (*Known* vs. *Unknown*), **Structure** (*Structured* vs. *Unstructured*), and **Disposition** (*Moral* vs. *Immoral*) under controlled, identical conditions.

### 4.1 Experiment Setup

**Benchmarks** To evaluate the role-playing performance of RPAs, we adopt two benchmarks: PersonaGym (Samuel et al., 2024) for single-turn interviewing, and CoSER (Wang et al., 2025c) for multi-turn interactions. PersonaGym evaluates response quality via dynamic, persona-tailored interviews using five rubrics: *Persona Consistency*, *Linguistic Habits*, *Expected Action*, *Action Justification*, and *Toxicity Control*. The averaged *Persona Score* serves as our default metric on PersonaGym.

Next, CoSER assesses multi-party narratives expanded from seed scenarios using three rubrics: *Anthropomorphism*, *Character Fidelity*, and *Storyline Quality*. The averaged *Persona Score* serves as our default metric on CoSER. Since our dataset does not contain scenario information, we use gemini-2.5-pro (Comanici et al., 2025) to generate scenarios specifically tailored to the participating characters to serve as narrative seeds. To reduce scenario-specific bias, we evaluate each work with three distinct seed scenarios; each session uses three characters from the same material and runs for 18 turns. For the **Disposition** axis, CoSER sessions are labeled as *Moral* when they include 0–1 immoral characters and *Immoral* when they include 2–3. We provide further details in Appendix D.

**Backbone LLMs** We employ five recent open-source LLMs as our backbone role-playing models: Qwen3-8B, Qwen3-235B-A22B-Instruct (Yang et al., 2025a), Mistral-Small-3.2-24B (Liu et al., 2026), OLMo-3.1-32B-Instruct (Olmo et al., 2025), and Deepseek-v3.2 (Liu et al., 2025). For quantitative evaluation on both benchmarks, we utilize GPT-4o (Hurst et al., 2024) as the Judge.

### 4.2 Main Results

Table 2 presents RPA performance across all three axes. The results reveal a clear asymmetry: *Familiarity* and *Structure* show negligible impact, while *Disposition* produces large and consistent performance degradation across all models. We further validate our main findings by human evaluation, which exhibits an identical trend (see Appendix E.1 for details). We also provide per-metric breakdowns in Appendix F.1.

**Familiarity: Task-dependent but not a stable driver.** The **Familiarity** axis exhibits a benchmark-dependent pattern rather than a stable advantage in either direction. On PersonaGym,

| | Familiarity | | | | Structure | | | | Disposition | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Known* | *Unknown* | Δ | ***p*** | *Str.* | *Unstr.* | Δ | ***p*** | *Moral* | *Immoral* | Δ | ***p*** |
| **PersonaGym (Single-turn interview)** | | | | | | | | | | | | |
| **Qwen3-8B** | 4.50 | 4.43 | -0.07 | * | 4.46 | 4.47 | 0.01 | - | 4.64 | 4.29 | **-0.35** | *** |
| **Qwen3-235B** | 4.66 | 4.62 | -0.04 | - | 4.64 | 4.61 | -0.03 | - | 4.8 | 4.47 | **-0.33** | *** |
| **Mistral-Small** | 4.58 | 4.53 | -0.05 | * | 4.55 | 4.54 | -0.01 | - | 4.71 | 4.39 | **-0.32** | *** |
| **Olmo3.1-32B** | 4.68 | 4.68 | 0.00 | - | 4.68 | 4.62 | -0.07 | - | 4.72 | 4.51 | **-0.20** | ** |
| **DeepSeek-v3.2** | 4.53 | 4.54 | 0.01 | - | 4.54 | 4.51 | -0.03 | - | 4.68 | 4.39 | **-0.29** | *** |
| **CoSER (Multi-turn interaction)** | | | | | | | | | | | | |
| **Qwen3-8B** | 20.67 | 23.44 | 2.77 | - | 22.06 | 22.37 | 0.31 | - | 24.88 | 18.88 | **-6.00** | *** |
| **Qwen3-235B** | 33.30 | 35.00 | 1.70 | - | 34.15 | 31.66 | -2.49 | ** | 38.06 | 29.66 | **-8.40** | *** |
| **Mistral-Small** | 31.30 | 33.30 | 2.00 | - | 32.30 | 33.84 | 1.54 | * | 35.04 | 29.15 | **-5.89** | *** |
| **Olmo3.1-32B** | 22.77 | 24.78 | 2.01 | - | 23.55 | 24.38 | 0.83 | - | 26.81 | 20.28 | **-6.53** | *** |
| **DeepSeek-v3.2** | 38.10 | 42.63 | 4.53 | * | 40.36 | 39.44 | -0.92 | - | 44.82 | 35.60 | **-9.22** | *** |

Table 2: RPA performance across three axes (Familiarity, Structure, Disposition) on PersonaGym and CoSER. denotes the difference between conditions; significance levels: * p<0.05, ** p<0.01, *** p<0.001.

*Known* characters obtain slightly higher scores for some models, but the effect sizes remain very small ($|\Delta| \leq 0.07$). On CoSER, *Unknown* characters show a higher-score tendency across models, but only DeepSeek-v3.2 reaches statistical significance ($p < 0.05$). We interpret this reversal as an interaction between canonical priors and contextual grounding: prior knowledge may help in single-turn interviews, while accumulated profile and scenario context in multi-turn narratives can dilute or compete with such priors. Appendix F.1 further supports this interpretation through a turn-wise analysis of the Familiarity axis.

**Structure: Marginal and inconsistent.** Profile **Structure** shows virtually no impact on performance. On PersonaGym, the gap between *Structured* and *Unstructured* profiles is at most 0.03 and never reaches statistical significance. On CoSER, while Qwen3-235B and Mistral-Small show marginal significance, the effect sizes remain small ($\Delta \leq 2.49$), and the direction is inconsistent. Structured formatting does not provide a meaningful advantage over free-form narratives.

**Disposition: Large and consistent degradation.** In stark contrast, **Disposition** axis produces the largest and most consistent performance gap across all conditions. On PersonaGym, *Immoral* characters score significantly lower than *Moral* ones across every model ($\Delta$: 0.20–0.35, all $p < 0.01$). On CoSER, the gap widens further ($\Delta$: 5.89–9.22, all $p < 0.001$), with DeepSeek-v3.2 exhibiting the largest drop of 9.22 points. Unlike the other two axes, disposition effect is uniformly significant, large in magnitude, and consistent in direction across both benchmarks and all models.

| Model | Familiarity | | Disposition | |
|---|---|---|---|---|
| | ***K*** | ***UK*** | ***M*** | ***IM*** |
| **PersonaGym** | | | | |
| Qwen3-235B | -0.17 | -0.18 | **-0.82** | **+0.26** |
| DeepSeek-v3.2 | -0.16 | -0.11 | **-0.62** | **+0.34** |
| **CoSER** | | | | |
| Qwen3-235B | -1.02 | +0.81 | **-22.17** | **+3.47** |
| DeepSeek-v3.2 | -0.55 | -0.21 | **-19.89** | **+1.79** |

Table 3: Counterfactual rewrite results. K and UK denote *Known* and *Unknown*, respectively, while M and IM denote *Moral* and *Immoral*.

### 4.3 Analyses & Discussion

**Counterfactual rewrite validation.** To examine whether the observed gaps follow the intended profile axes rather than fixed character identities, we conduct counterfactual rewrite experiments on shared base profiles. For the **Familiarity** axis, we swap the `Personal Attributes` field between paired *Known* and *Unknown* profiles; for the **Disposition** axis, we rewrite the remaining fields toward the opposite moral polarity. Table 3 reports the score change from the original to the counterfactual profile. The Familiarity swap yields negligible changes, within $\pm 0.2$ on PersonaGym and $\pm 1.1$ on CoSER. In contrast, the Disposition rewrite produces large directional shifts: *Moral*→*Immoral* sharply degrades performance, up to $-0.82/-22.17$ on PersonaGym/-CoSER, while *Immoral*→*Moral* improves it, up to $+0.34/+3.47$. Together, these results support our main finding that **Familiarity** is not a major driver, while the **Disposition** gap follows dispositional content rather than character identity. Detailed procedure and results are provided in Appendix F.2.

| Rubrics | *M* | *IM* | Δ | *p* |
|---|---|---|---|---|
| **Instruct-SFT** | | | | |
| *Anthro.* | 10.78 | 6.02 | -4.76 | * |
| *Char. Fid.* | 15.88 | 13.44 | -2.44 | - |
| *Story Qual.* | 32.45 | 26.66 | -5.79 | * |
| *Avg.* | 19.70 | 15.37 | -4.53 | * |
| **+ DPO** | | | | |
| *Anthro.* | 13.53 | 6.32 | **-7.21** | *** |
| *Char. Fid.* | 16.67 | 7.01 | **-9.66** | *** |
| *Story Qual.* | 46.08 | 39.95 | -6.13 | ** |
| *Avg.* | 25.42 | 17.79 | **-7.63** | *** |

Table 4: Alignment-stage ablation on CoSER. *M*, *IM* denote *Moral*, *Immoral*; *Anthro.*, *Char. Fid.*, and *Story Qual.* denote *Anthropomorphism*, *Character Fidelity*, and *Storyline Quality*. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

**Alignment-stage ablation.** Because the **Disposition** axis is conceptually tied to post-SFT alignment, we further examine whether alignment amplifies the Moral–Immoral asymmetry. To this end, we compare two checkpoints from the same model family: OLMo-3.1-32B-Instruct-SFT before preference alignment and OLMo-3.1-32B-Instruct-DPO after DPO alignment (Rafailov et al., 2023). As shown in Table 4, DPO improves the average CoSER scores for both *Moral* and *Immoral* characters, but the improvement is larger for *Moral* characters, widening the average Moral–Immoral gap from $-4.53$ to $-7.63$. This widening is most pronounced in *Character Fidelity*: the gap is small and not significant under Instruct-SFT ($\Delta = -2.44$), but becomes large and significant after DPO ($\Delta = -9.66$). Notably, while DPO slightly improves *Character Fidelity* for *Moral* characters, it substantially decreases it for *Immoral* characters. These results suggest that DPO-style preference alignment, which favors human-preferred helpful or benign responses, can selectively weaken fidelity to immoral profile content. We further evaluate a subsequent RLVR-aligned checkpoint (Lambert et al., 2025) and find that it partially improves overall role-playing scores, but does not close the DPO-amplified Moral–Immoral gap; see Appendix F.3.

**Field-level localization of the Disposition bottleneck.** Building on the finding that the **Disposition** axis drives the largest performance disparities, we further examine whether this degradation is holistic or localized to specific profile dimensions. Using the field-categorized character set (Section 3.2), we track CoSER performance as the number of *Immoral* characters increases.

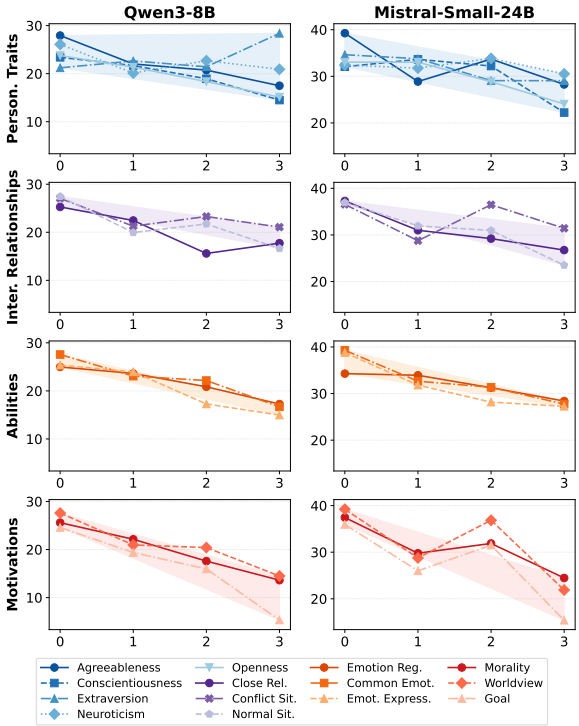


Figure 3: Field-level localization of the Disposition bottleneck on CoSER. X-axis denotes the number of *Immoral* characters included in each sample, and the Y-axis denotes the *CoSER Score*.

As shown in Figure 3, the degradation is not uniform across fields. `Personality Traits` remain relatively stable, suggesting that models can still reproduce surface-level dispositional signals such as extraversion or agreeableness. In contrast, fields in `Motivations` exhibit the steepest and most consistent decline, especially in value-laden fields such as *Morality* and *Goal*. Fields in `Abilities` and `Interpersonal Relationships` show intermediate degradation. These results indicate that the primary bottleneck lies not in adopting basic character traits, but in faithfully representing specific value-laden profile fields under *Immoral* settings. This motivates the targeted mitigation strategy introduced in Section 5. Full localization results across all five models are provided in Appendix F.4.

## 5 Mitigating the Alignment Bottleneck

Our analysis shows that this bottleneck is concentrated in specific value-laden fields, including `Motivations` fields such as *Goal* and *Morality*, as well as disposition-sensitive fields in other profile dimensions. Now, our objective is to improve the fidelity of *Immoral* characters without compromising the baseline performance of *Moral* ones. To this end, we introduce *Field-Aware Contrastive Decod-*

*ing* (FACD), a training-free strategy that contrasts the full profile against a sanitized profile in which immoral or disposition-sensitive fields are removed, thereby selectively amplifying the suppressed profile signals at inference time. We further validate our results by human evaluation, which exhibits an identical trend (see Appendix E.2 for details).

### 5.1 Preliminary: Contrastive Decoding

Modern alignment procedures instill strong helpful-assistant priors into LLMs (Ouyang et al., 2022; Huang et al., 2025), which persistently resist conflicting behavioral instructions delivered through system prompts (Sharma et al., 2024). *Contrastive Decoding* (Li et al., 2023b) offers a principled remedy by reformulating generation as a contrast between two output distributions, thereby isolating and amplifying a desired behavioral signal at decoding time without any parameter updates.

Formally, consider an autoregressive language model $M_\theta$ that produces next-token logits $z_t = \log p_\theta(x_t \mid \mathbf{x}_{<t})$ given a prefix $\mathbf{x}_{<t}$. (Dong et al., 2026) adapts the contrastive principle to system-prompt adherence by running two parallel forward passes with different system prompts: a *positive prompt* $s_{\text{pos}}$ encoding the desired persona, and a *negative prompt* $s_{\text{neg}}$ representing the behavior we wish to suppress. Let $z_t^{\text{pos}}$ and $z_t^{\text{neg}}$ denote the logits produced under each prompt, respectively. The steered logits are then defined as:

$$\tilde{z}_t = z_t^{\text{pos}} + \alpha\,(z_t^{\text{pos}} - z_t^{\text{neg}}), \qquad (1)$$

where $\alpha$ controls the amplification strength. At $\alpha$=0, generation reduces to standard decoding under $s_{\text{pos}}$. As $\alpha$ increases, the contrast term $z_t^{\text{pos}} - z_t^{\text{neg}}$ is amplified, suppressing tokens favored by the sanitized baseline ($s_{\text{neg}}$) while boosting tokens specific to the desired persona ($s_{\text{pos}}$).

### 5.2 Field-Aware Contrastive Decoding

**Motivation.** While standard contrastive decoding uniformly amplifies the entire persona signal, the bottleneck identified in Section 4 requires a more targeted intervention. When a profile prescribes adversarial traits, the alignment prior actively suppresses the specific tokens needed to portray them. Crucially, this suppression is not uniform across the profile: fields like `Goal` suffer steep declines, whereas polarity-insensitive fields like `Extraversion` remain stable (Figure 3). Therefore, we propose a *field-aware* variant that directs contrastive amplification exclusively toward the immoral signals that alignment suppresses.

**Method.** Let $\mathcal{F}$ denote the set of all fields in a character profile, and let $s(\mathcal{F}')$ denote the full role-playing prompt whose character profile is restricted to a field subset $\mathcal{F}' \subseteq \mathcal{F}$. The positive prompt $s_{\text{pos}} = s(\mathcal{F})$ uses the complete profile.

FACD constructs the negative prompt as a sanitized profile baseline. Starting from the full profile $\mathcal{F}$, we remove all non-Personal-Attributes fields classified as immoral by MoralBERT (Preniqi et al., 2024). Let $\mathcal{F}_{\text{PA}}$ denote the Personal Attributes fields and $\mathcal{F}_{\text{immoral}} = \{f \in \mathcal{F} \setminus \mathcal{F}_{\text{PA}} : M(f) = \text{immoral}\}$, where $M$ is MoralBERT. The retained fields are $\mathcal{F}_{\text{base}} = \mathcal{F} \setminus \mathcal{F}_{\text{immoral}}$.

If fewer than 6 non-PA fields remain after filtering, we add a fallback set $\mathcal{F}_{\text{pad}}$ drawn from disposition-robust fields identified in Figure 3, such as `Extraversion`, `Neuroticism`, and `Openness`. The final negative prompt is $s_{\text{neg}} = s\big(\mathcal{F}_{\text{base}} \cup \mathcal{F}_{\text{pad}}\big)$. Thus, $s_{\text{neg}}$ is a sanitized version of the original profile produced after both immoral-field filtering and disposition-robust padding. As a result of this full construction pipeline, the final $s_{\text{neg}}$ contains no field that is both disposition-sensitive and immoral. The steered logits follow Eq. 1; the contrast $z_t^{\text{pos}} - z_t^{\text{neg}}$ therefore amplifies the profile signals contributed by the omitted immoral fields, especially when they occur in disposition-sensitive parts of the profile. We set $\alpha$=1 as a moderate value that empirically balances fidelity gains and narrative coherence. We provide more details about FACD in Appendix H.

### 5.3 Mitigation Results

**FACD narrows the *Moral–Immoral* gap.** As shown in the top table of Figure 4, the default setting yields a substantial performance gap between *Moral* (M) and *Immoral* (IM): $-6.00$ for Qwen3-8B and $-5.89$ for Mistral-Small on average. Applying FACD reduces these gaps to $-4.50$ and $-2.22$, respectively. Importantly, this improvement does not come from sacrificing *Moral* performance for *Immoral* recovery: the *Moral* average increases for Qwen3-8B ($24.88 \rightarrow 27.07$) and remains comparable for Mistral-Small ($35.04 \rightarrow 34.31$). These results indicate that FACD mitigates the dispositional asymmetry while preserving overall role-playing quality, and can even benefit *Moral* settings when value-laden profile fields are under-expressed under default decoding.

| Dim. | Qwen3-8B | | | | Mistral-Small | | | |
|---|---|---|---|---|---|---|---|---|
| | M | IM | Δ | *Avg.* | M | IM | Δ | *Avg.* |
| **Default** | | | | | | | | |
| *Anthro.* | 14.53 | 8.55 | -5.98 | 11.72 | 20.27 | 17.85 | -2.42 | 19.13 |
| *Char. Fid.* | 11.39 | 6.39 | -5.00 | 9.04 | 27.12 | 15.78 | -11.34 | 21.78 |
| *Story Qual.* | 48.72 | 41.71 | -7.01 | 45.42 | 57.74 | 53.82 | -3.92 | 55.90 |
| *Avg.* | 24.88 | 18.88 | -6.00 | 22.06 | **35.04** | 29.15 | -5.89 | 32.27 |
| **Field-aware Contrastive Decoding (FACD)** | | | | | | | | |
| *Anthro.* | **18.24** | **11.39** | -6.85 | **15.02** | **21.01** | **18.37** | -2.64 | **19.77** |
| *Char. Fid.* | **13.91** | **14.38** | **+0.47** | **14.13** | 26.22 | **24.38** | **-1.84** | **25.35** |
| *Story Qual.* | **49.06** | **41.95** | -7.11 | **45.71** | 55.71 | 53.52 | **-2.19** | 54.68 |
| *Avg.* | **27.07** | **22.57** | **-4.50** | **24.95** | 34.31 | **32.09** | **-2.22** | **33.27** |

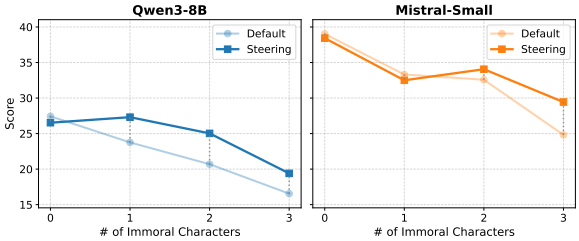


Figure 4: Comparison of steering performance and mitigation results across models. Y-axis of the lower figure indicates the CoSER score.

**Character Fidelity shows targeted recovery.** The largest recovery appears in *Character Fidelity*, the metric most directly tied to adherence to the given profile. This matches our diagnosis that the bottleneck primarily concerns the under-expression of immoral content in the disposition-sensitive profile field. On Qwen3-8B, the IM Character Fidelity score more than doubles from 6.39 to 14.38, closing the gap from $-5.00$ to $+0.47$. On Mistral-Small, it rises from 15.78 to 24.38, reducing the gap from $-11.34$ to $-1.84$. Meanwhile, *Anthropomorphism* and *Storyline Quality* largely preserved on both models. Thus, FACD does not merely make generations more intense or broadly persona-like; it primarily recovers profile-grounded traits that are otherwise under-expressed.

**Effective under stronger dispositional pressure.** The lower panel of Figure 4 shows that FACD is beneficial as the number of *Immoral* characters increases. Although the downward trend is not fully eliminated, FACD consistently lifts performance above the default baseline, with the largest gains appearing when all three characters are *Immoral*: Qwen3-8B improves from roughly 16 to 20, and Mistral-Small from 25 to 29. This indicates that FACD is most valuable in challenging multi-character settings, where the need to jointly instantiate multiple immoral personas makes the dispositional bottleneck more severe.

| Negative Prompt | M | IM | Δ | *Avg.* |
|---|---|---|---|---|
| *Personal Attributes only* | 22.02 | 18.40 | -3.62 | 20.21 |
| *w/o. Personality Traits* | 23.27 | 18.81 | -4.46 | 21.04 |
| *w/o. Interpersonal Relationships* | 23.18 | 20.53 | -2.65 | 21.86 |
| *w/o. Motivations* | 25.64 | 20.42 | -5.22 | 23.03 |
| *w/o. Abilities* | 24.61 | 19.60 | -5.01 | 22.11 |
| *w/o. Value-laden fields* | 24.75 | 21.11 | -3.64 | 22.93 |
| ***FACD*** | **27.07** | **22.57** | -4.50 | **24.82** |

Table 5: Ablation study on the negative prompt construction for FACD. *M* and *IM* denote *Moral* and *Immoral*, respectively. Δ denotes the Moral–Immoral gap, and *Avg.* denotes the mean of $M$ and $IM$.

## 5.4 Ablation Study

We ablate the construction of the FACD negative prompt by retaining only *Personal Attributes*, removing each top-level field group, or removing all disposition-sensitive fields regardless of Moral or Immoral polarity. As shown in Table 5, the *Personal Attributes only* baseline yields low scores for both groups, suggesting that an overly sparse negative prompt fails to provide a sufficiently character-grounded contrast. Among the field-removal variants, excluding more disposition-sensitive dimensions such as *Motivations* and *Abilities* is more effective than excluding relatively surface-level *Personality Traits*. Although some ablations produce a smaller Moral–Immoral gap, this mainly reflects degraded *Moral* performance rather than improved *Immoral* recovery, making such reductions less practical. Removing all value-laden fields also performs competitively, but still falls short of FACD. This supports the core design of FACD: an effective negative prompt should be both character-grounded and polarity-aware, preserving enough profile context while selectively omitting immoral/disposition-sensitive fields. This allows the contrastive term to amplify profile-specific signals from the full profile rather than generic assistant behavior.

## 6 Conclusion

We present a diagnostic framework that disentangles role-playing fidelity along three axes—Familiarity, Structure, and Disposition—revealing that moral Disposition is the primary bottleneck, driven by alignment priors suppressing tokens needed for immoral portrayal. To address this, we propose Field-Aware Contrastive Decoding, a training-free strategy that selectively amplifies suppressed traits, significantly narrowing the Moral–Immoral gap without fine-tuning.

## Limitations

While our study provides a foundational disentanglement of character profile axes, it has two limitations that offer avenues for future work. First, we adopted a binary classification of **Disposition**. Although this simplified categorization was essential for our diagnostic study to clearly identify the causal impact of alignment priors on role-playing fidelity, it may not fully capture the complexity of "anti-heroes" or morally ambiguous personas found in nuanced narratives. Second, due to computational resource constraints inherent in running parallel forward passes for contrastive decoding, our proposed Field-Aware Contrastive Decoding (FACD) was evaluated on a subset of the frontier LLMs. Despite these constraints, the consistent performance gains observed in the evaluated models underscore FACD's potential as a robust, training-free mitigation strategy, providing a strong basis for future scaling to even larger model architectures.

## Ethics Statement

This paper compares role-playing behavior under two sources of character information: (i) Known character profiles derived from publicly accessible, community-curated Fandom pages, and (ii) Unknown character profiles generated by an LLM (gpt-oss). We provide clear attribution for third-party sources and document our generation procedure; we release only the minimum derived, structured metadata needed for research rather than reproducing copyrighted source text verbatim, and we expect downstream users to comply with applicable licenses/terms. Because both fan-curated metadata and model generation may contain or imply harmful content (e.g., violence, harassment, stereotyping), we apply filtering and normalization using a safety-aligned Claude model and conduct manual spot checks, while noting that residual risks may remain. Our evaluation instantiates these profiles within CoSER (Wang et al., 2025c) and PersonaGym (Samuel et al., 2024) pipelines and follows their ethical framing: persona/role-playing systems can be misused to generate targeted harmful content, reinforce stereotypes, or encourage anthropomorphization, and persona construction may raise intellectual-property and privacy concerns (especially if adapted to real individuals), which would require explicit consent and stronger protections. Finally, since we use LLMs for generation and/or automated assessment, we acknowledge potential judge/model biases and report configurations to support reproducibility, and we do not position these methods for safety-critical deployment.

## References


Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of ACL 2020*, pages 4190–4197.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Arney Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2025. Efficient llm inference via chunked prefills. *SIGOPS Oper. Syst. Rev.*, 59(1):9–16.

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, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2025. Introducing claude sonnet 4.5.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2024. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*.

Sijia Cheng, Wen Yu Chang, and Yun-Nung Chen. 2025. Exploring personality-aware interactions in salesperson dialogue agents. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 60–71, Bilbao, Spain. Association for Computational Linguistics.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. 2024. Examining identity drift in conversations of llm agents. *arXiv preprint arXiv:2412.00804*.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan

Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3326–3346.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, and Nigel Collier. 2026. Steer model beyond assistant: Controlling system prompt strength via contrastive decoding. *arXiv preprint arXiv:2601.06403*.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 120–127. IEEE.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

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

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT 2019*, pages 3543–3556.

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

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

Yonghyun Jun and Hwanhee Lee. 2025. Exploring persona sentiment sensitivity in personalized dialogue generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18384–18402, Vienna, Austria. Association for Computational Linguistics.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, et al. 2024. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. *Advances in Neural Information Processing Systems*, 37:49403–49428.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Dan P McAdams and Jennifer L Pals. 2006. A new big five: fundamental principles for an integrative science of personality. *American psychologist*, 61(3):204.

Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: a fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 433–442.

Cheng Qian, Xinran Zhao, and Tongshuang Wu. 2024. "merge conflicts!"' exploring the impacts of external knowledge distractors to parametric knowledge graphs. In *First Conference on Language Modeling*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479, Miami, Florida, USA. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 2023. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, pages 31094–31116. PMLR.

Libo Sun, Siyuan Wang, and Zhongyu Wei. 2025. Identity-driven hierarchical role-playing agents. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 403–417. Springer.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL 2019*, pages 4593–4601.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao,

Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024b. 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.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025b. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.

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. 2025c. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024c. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From personas to talks: Revisiting the impact of personas on LLM-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5439–5453, Suzhou, China. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

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

Bohao Yang, Dong Liu, Chenghao Xiao, Kun Zhao, Chen Tang, Chao Li, Lin Yuan, Yang Guang, and Chenghua Lin. 2025b. Crafting customisable characters with LLMs: A persona-driven role-playing agent framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20216–20240, Suzhou, China. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. 2025. Revealing and mitigating the challenge of detecting character knowledge errors in LLM role-playing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33267–33290, Suzhou, China. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, et al. 2024. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. 2025. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

# A Appendix Table of Contents

We organize the Appendix as follows:

1. Section B provides the overall usage of Large Language Models (LLMs).
2. Section C provides dataset details, including the full character-profile schema, known and unknown character construction pipelines, moral–immoral disposition validation, unstructured-profile generation, and dataset validation.
3. Section D describes the experimental setup, including model and API configurations, benchmark protocols for PERSONAGYM and COSER, scenario generation, and implementation details for the evaluation pipeline.
4. Section E reports human evaluation results for both RQ1 and RQ2, validating the automatic evaluation trends for the main axis analysis and the FACD mitigation results.
5. Section F provides detailed RQ1 analyses, including per-metric results for the Structure, Familiarity, and Disposition axes, the counterfactual rewrite prompt, and the full field-level localization results across models.
6. Section G further analyzes the Familiarity axis through turn-length ablations and attention/saturation diagnostics, showing how multi-turn interaction changes reliance on profile, history, and self-generated context.
7. Section H gives additional details on Field-Aware Contrastive Decoding, including field-level immoral filtering, a concrete Moriarty example, and the disposition-robust fallback fields used to construct the sanitized negative prompt.

# B Use of Large Language Models

We used large language models as assistance tools in several parts of the research process. First, we used ChatGPT for improving the clarity of the manuscript; all substantive claims, analyses, and final wording were reviewed and edited by the authors. Second, we used LLMs (Claude-sonnet-4.5, Claude-haiku-4.5, Gemini-2.5-pro) during the construction and processing of character profiles, including [profile generation / summarization / rewriting / filtering]. Third, we used GPT-4o as LLM-as-a-Judge for evaluating role-playing performance. We use same prompt templates described in prior works: PersonaGym (Samuel et al., 2024) and CoSER (Wang et al., 2025c). We also manually inspected the outputs. Finally, we used Claude Code as a coding assistant for implementation support, including code drafting, debugging, and refactoring. All generated or modified code was reviewed, tested, and verified by the authors.

# C Dataset Details

## C.1 Dataset Schema

The dataset is structured using a hierarchical schema that represents character identity beyond personality traits alone, decomposing it into multiple high-level dimensions such as personal attributes, personality traits, relationships, motivations, and abilities. Detailed field definitions and example schema templates are provided in Table 13 and Table 14. For profile generation prompt, see Table 15.

## C.2 Dataset Construction Details

### C.2.1 Known Character Profiles

We constructed character profiles for known characters using character metadata from Fandom. For each work, we extracted 3–5 characters, restricting our selection to widely recognized franchises and prioritizing high-salience characters with substantial narrative prominence. The resulting pool of selected works and the number of characters chosen per work are summarized in Table 16.

### C.2.2 Unknown Character Profiles

Unknown characters are constructed through a structured prompt-based pipeline that mirrors the schema and construction procedure used for known characters. The pipeline consists of three components.

First, we define a profile skeleton that specifies the demographic attributes, personality trait seeds, and scenario genres from which unknown characters are instantiated. This skeleton provides only structural constraints and candidate value pools, without enforcing narrative coherence. The full schema and organization are shown in Table 17.

Second, to ensure internal logical consistency among sampled attributes, we employ an LLM-based profile validation prompt. This judge prompt evaluates whether a sampled profile contains intrinsic conflicts (e.g., age–education–occupation or family status–age mismatches) and assigns a coherence score on a 1–10 scale. Only profiles that

satisfy basic logical coherence are retained. The validation prompt is detailed in Table 18.

Finally, for downstream experiments requiring narrative context, we use a separate story generation prompt that conditions on the validated profile to produce short descriptive scenes. This prompt is used solely to generate auxiliary narrative material and does not alter the underlying character attributes. The corresponding prompt template is provided in Table 19.

#### C.2.3 Verifying Disposition Classification with Human Validation

| Fields | Agreement | $\kappa$ |
|---|---|---|
| `Personality Traits` | 0.875 | 0.744 |
| `Interpersonal Relationships` | 0.9 | 0.8 |
| `Abilities` | 0.85 | 0.70 |
| `Motivations` | 0.95 | 0.89 |
| *Overall Profile* | **0.90** | **0.80** |

Table 6: LLM–human agreement and Cohen's $\kappa$ scores for Moral–Immoral disposition classification.

To validate the *Moral-Immoral* classification, we sample 20 *Moral* and 20 *Immoral* profiles and have three graduate students in an AI research lab with documented English proficiency annotators independently label them, determining the final label through majority voting. As shown in Table 6, the agreement between the LLM and human classifications for the overall profiles reaches 0.90 with a Cohen's $\kappa$ of 0.80, confirming reliable disposition labeling across all profile dimensions. Furthermore, the agreement scores for the individual detailed fields consistently exceed 0.85, with corresponding $\kappa$ values all above 0.70, demonstrating substantial inter-annotator consistency. Also, we provide full classification prompt in Table 21.

#### C.2.4 Unstructured Profiles

Based on the structured profiles we developed, we generated a parallel set of unstructured profiles. The detailed prompt used for this process is provided in Table 22.

### C.3 Dataset Validation

**Field Independence** To verify each profile field captures distinct information, we compute Gemini-001 embeddings (Lee et al., 2025) for all leaf fields and visualize them via t-SNE (Maaten and Hinton, 2008). As shown in Figure 5, fields cluster separately by dimension, confirming that the schema produces non-overlapping content across fields.

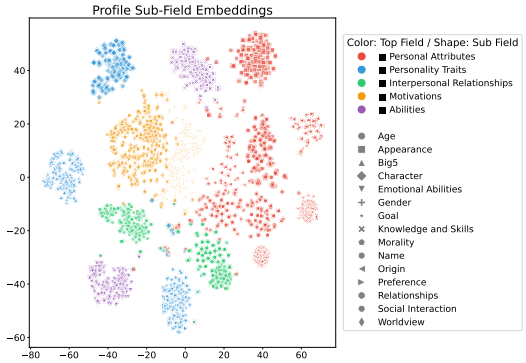


Figure 5: t-SNE visualization of profile sub-field embeddings, colored by dimension and shaped by leaf field.

**Profile Reliability** To ensure the profiles accurately reflect the metadata without hallucination, we use GPT-4o to decompose each profile into atomic facts (Sadeq et al., 2024). We then segment the metadata into chunks of up to 1,200 words. For each atomic fact, we retrieve the top-3 most semantically similar chunks and prompt Gemini-3.1-Pro to verify whether these evidence chunks support the given fact. Our analysis reveals that the profiles are decomposed into an average of 153.3 atomic facts, and 141.04 of these are explicitly supported by the metadata chunks, achieving a precision score of 0.92. This confirms that our profiles are highly faithful to the source material and possess robust reliability.

## D Experimental Details

Most models were accessed via the OpenRouter API with temperature set to 0 (greedy decoding) for role-playing agents and judge models. We used the following providers: TogetherAI for Qwen3 and GPT-oss families, and AtlasCloud for the Deepseek-v3.2 model. For mechanical analysis, we used 2 NVIDIA A6000 GPUs.

### D.1 Benchmark Details

**PersonaGym** PERSONAGYM evaluates persona agents using a deliberately *single-character*, *single-turn protocol*: each evaluation instance assigns exactly one persona (typically via a persona-conditioning system instruction) and asks one task question, and the agent produces one response; instances are scored independently, so the benchmark does not test multi-turn state tracking. In the full framework, an LLM "reasoner" first selects persona-relevant contexts from a pool of 150 environments, then generates task-specific questions for the chosen settings; the released static

benchmark fixes this process into 200 personas and 10,000 questions for standardized comparison across models. PERSONAGYM covers five task dimensions—Action Justification, Expected Action, Linguistic Habits, Persona Consistency, and Toxicity Control—and scores each single-turn response with detailed, task-specific rubrics on a 1–5 ordinal scale, augmented with persona-and-question–conditioned exemplar responses to calibrate judging. Each response is graded by an ensemble of two evaluator LLMs, and the final score is computed by averaging the judges' outputs; results are then aggregated across tasks into an overall *PersonaScore*.

**CoSER** COSER evaluates role-playing language agents in multi-turn, multi-character literary scenes via Given-Circumstance Acting (GCA), where an "actor" LLM is required to sequentially portray multiple characters to reconstruct authentic scenarios extracted from well-known novels. Concretely, COSER provides multi-party dialogues and rich contextual conditioning signals (e.g., scenario descriptions, character profiles, motivations, and—in the dataset representation—optional speech/action/thought annotations), and explicit support for >2-character interactions. At evaluation time, GCA first runs a multi-agent simulation: one RPA instance is created per character using the same actor LLM, each conditioned on the shared scenario plus its own persona materials, while a next-speaker-prediction (NSP) component selects who speaks next and a separate "environment" model emits environmental feedback; the rollout terminates either when NSP outputs an END signal or when a maximum turn budget (e.g., 18 turns) is reached. To increase controllability, GCA additionally applies penalty-based LLM judging: instead of asking a judge to output a single holistic score, a critic LLM identifies rubric-defined flaws (e.g., deviations from the reference dialogue or lack of initiative), assigns each flaw a severity level (1–5), aggregates penalties into per-dimension scores, and applies a length-correction procedure to mitigate bias toward conversation length. The rubric-driven scores are reported over three core dimensions: Anthropomorphism (the extent to which RPAs speak and act in a human-like manner), Character Fidelity (the degree to which the character profile is faithfully reflected in the generated narrative), and Storyline Quality (how natural and engaging the narrative flow is). In our setting, because our dataset instances do not provide explicit scenario descriptions, we generate synthetic scenarios using Gemini-2.5-Pro (temperature = 0.7), conditioning on the set of characters participating in each conversation. To mitigate potential bias induced by any single scenario formulation, we construct three clearly distinct seed scenarios per work and use these as the basis for interaction. For scenario generation prompt, see Table 20.

Notably, both benchmarks have demonstrated strong correlation with human judgment in their original studies.

# E Human Evaluation

We conduct human evaluation on both PersonaGym and CoSER with two backbone models (Qwen3-235B and Mistral-Small). For each benchmark, we randomly sample 20% of characters from each condition and divide them into three non-overlapping groups. Three annotators with bachelor's degrees and English proficiency each evaluate one group across both models. Results are reported in Table 7. The same character–model pairs are reused for the FACD mitigation evaluation (Table 8).

## E.1 RQ1 Human Validation Results

| | Familiarity | | Disposition | |
|---|---|---|---|---|
| | Known | Unknown | Moral | Immoral |
| **PersonaGym (Single-turn interview)** | | | | |
| **Qwen3-8B** | 4.29 | 4.25 | 4.32 | 4.11 |
| **Mistral-Small** | 4.20 | 4.16 | 4.34 | 4.21 |
| **CoSER (Multi-turn interaction)** | | | | |
| **Qwen3-8B** | 3.00 | 2.89 | 3.19 | 2.69 |
| **Mistral-Small** | 3.47 | 3.53 | 3.63 | 3.39 |

Table 7: Human evaluation scores for RQ1 across Familiarity and Disposition axes on both benchmarks.

As shown in Table 7, human judgments are consistent with the LLM-based evaluation: the Disposition axis produces the largest performance gap, while Familiarity differences remain marginal.

## E.2 RQ2 Human Validation Results

Table 8 confirms that FACD improves Immoral character scores under human evaluation as well, aligning with the automatic evaluation trends reported in Section 5.3.

# F RQ1 Detailed Analysis

This section provides detailed per-metric results across all five models, further analyses, and used prompts examined in Section 4.

| | Moral | Immoral | *Avg.* |
|---|---|---|---|
| **Default** | | | |
| **Qwen3-8B** | 3.19 | 2.69 | 2.94 |
| **Mistral-Small** | 3.63 | 3.39 | 3.51 |
| **Field-Aware Contrastive Decoding** | | | |
| **Qwen3-8B** | 3.56 | 3.25 | 3.40 |
| **Mistral-Small** | 3.68 | 3.47 | 3.58 |

Table 8: Human evaluation scores for FACD mitigation on CoSER.

| Model | Type | Familiarity *K* | Familiarity *UK* | Disposition *M* | Disposition *IM* |
|---|---|---|---|---|---|
| **PersonaGym** | | | | | |
| Qwen3-235B | *Original* | 4.66 | 4.62 | 4.80 | 4.47 |
| | *Counter* | 4.49 | 4.44 | 3.98 | 4.73 |
| | Δ | -0.17 | -0.18 | **-0.82** | **+0.26** |
| DeepSeek-v3.2 | *Original* | 4.53 | 4.54 | 4.68 | 4.39 |
| | *Counter* | 4.37 | 4.43 | 4.06 | 4.73 |
| | Δ | -0.16 | -0.11 | **-0.62** | **+0.34** |
| **CoSER** | | | | | |
| Qwen3-235B | *Original* | 33.30 | 35.00 | 38.06 | 29.66 |
| | *Counter* | 32.28 | 35.81 | 15.89 | 33.13 |
| | Δ | -1.02 | +0.81 | **-22.17** | **+3.47** |
| DeepSeek-v3.2 | *Original* | 38.10 | 42.63 | 44.82 | 35.60 |
| | *Counter* | 37.55 | 42.42 | 24.93 | 37.39 |
| | Δ | -0.55 | -0.21 | **-19.89** | **+1.79** |

Table 9: Counterfactual rewrite results. K and UK denote *Known* and *Unknown*, respectively, while M and IM denote *Moral* and *Immoral*.

### F.1 Detailed Results per Axis

**Structure Axis** Per-metric breakdowns for the *Structure* axis comparison are provided in Table 24.

**Familiarity Axis** Per-metric breakdowns for the *Familiarity* axis comparison are provided in Table 23.

**Disposition Axis** Per-metric breakdowns for the *Disposition* axis comparison are provided in Table 25.

### F.2 Details for Counterfactual Rewriting

To examine whether the observed gaps follow the intended profile axes rather than fixed character identities, we conduct counterfactual rewrite experiments on shared base profiles. For the **Familiarity** axis, we pair each *Known* profile with its embedding-nearest *Unknown* profile and swap only the `Personal Attributes` field. The resulting counterfactual profile shares the same `Personal Attributes` while all remaining fields come from the paired profile. As shown in Table 9, performance remains broadly stable after this identity swap. Although counterfactual profiles show a slight overall decline, likely due to conflicts between swapped identity attributes and the remaining profile content, the magnitude of this decline remains limited. This further supports our finding that **Familiarity** is not a major bottleneck in role-playing performance.

For the **Disposition** axis, we rewrite all non-Personal-Attributes fields of each *Moral* profile toward immoral content, and vice versa, using Claude-4.5-Haiku while preserving profile length, specificity, and writing style (full prompt is provided in Table 26). Unlike Familiarity, this rewrite produces clear directional shifts: *Moral*→*Immoral* substantially lowers performance on both PersonaGym ($-0.82/-0.62$) and CoSER ($-22.17/-19.89$), whereas *Immoral*→*Moral* improves performance on PersonaGym ($+0.26/+0.34$) and CoSER ($+3.47/+1.79$). These results provide additional evidence that the Disposition gap follows the rewritten dispositional content rather than the original character identity.

| Rubrics | *M* | *IM* | Δ | *p* |
|---|---|---|---|---|
| **Instruct-SFT** | | | | |
| *Anthro.* | 10.78 | 6.02 | -4.76 | * |
| *Char. Fid.* | 15.88 | 13.44 | -2.44 | - |
| *Story Qual.* | 32.45 | 26.66 | -5.79 | * |
| *Avg.* | 19.70 | 15.37 | -4.53 | * |
| **+ DPO** | | | | |
| *Anthro.* | 13.53 | 6.32 | **-7.21** | *** |
| *Char. Fid.* | 16.67 | 7.01 | **-9.66** | *** |
| *Story Qual.* | 46.08 | 39.95 | -6.13 | ** |
| *Avg.* | 25.42 | 17.79 | **-7.63** | *** |
| **+ RLVR** | | | | |
| *Anthro.* | 15.47 | 10.56 | -4.91 | * |
| *Char. Fid.* | 15.24 | 7.58 | -7.66 | *** |
| *Story Qual.* | 49.71 | 42.72 | **-6.99** | ** |
| *Avg.* | 26.81 | 20.28 | -6.53 | *** |

Table 10: Alignment-stage ablation on CoSER. *M*, *IM* denote *Moral*, *Immoral*; *Anthro.*, *Char. Fid.*, and *Story Qual.* denote *Anthropomorphism*, *Character Fidelity*, and *Storyline Quality*. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

### F.3 Additional RLVR-aligned checkpoint.

As discussed in Section 4.3, DPO alignment widens the Moral–Immoral gap from $-4.53$ to $-7.63$, largely by degrading *Immoral Character Fidelity*. We further evaluate OLMo-3.1-32B-Instruct after RLVR alignment, which optimizes verifiable objectives such as math and coding (Lambert et al.,

2025). As shown in Table 10, RLVR improves the average scores for both *Moral* and *Immoral* characters, suggesting incidental transfer of general reasoning or instruction-following ability to role-playing. However, it does not eliminate the DPO-amplified asymmetry: the average gap remains large and significant ($\Delta = -6.53$), and the *Character Fidelity* gap remains much larger than under Instruct-SFT ($-7.66$ vs. $-2.44$).

### F.4 Detailed Localization Results

Figure 7 provides the full five-model version of the localization analysis shown in Figure 3.

## G Detailed Analysis for Familiarity Axis

### G.1 Turn Ablation

#### G.1.1 Motivation and Setting

Interacting environments differ from interviewing settings in agent count (single vs. multi) and duration (single vs. multi-turn). Inspired by findings that parametric knowledge can be detrimental in long-term contexts (Qian et al., 2024; Xu et al., 2024), we hypothesize that the multi-turn factor suppresses the benefits of parametric identities, thereby bridging the gap between known and unknown characters. To test this, we track performance trends across varying turn lengths (6, 12, 18, 24) using diverse backbone models (Qwen3-8B, Qwen3-235B, Deepseek-v3.2) on the CoSER benchmark.

#### G.1.2 Preliminaries

To diagnose how a multi-turn role-playing agent reallocates reliance across competing context sources as the interaction horizon grows, we extract decoding-time *attention/activation* signals rather than relying solely on end-task performance curves (Jain and Wallace, 2019; Chefer et al., 2021). Concretely, for each generation, we partition the available prompt context into disjoint token segments (e.g., PROFILE, HISTORY, and GENERATED) and analyze the model's attention at the *final generated token*. Let $L$ be the number of transformer layers and $H$ the number of attention heads. At layer $\{\ell \in 1, \ldots, L\}$, we take the attention distribution from the last query position, average it across heads, and obtain $\bar{\alpha}^{(\ell)} \in [0,1]^T$ over the $T$ key/value positions. For a segment $s \subseteq \{1, \ldots, T\}$ with token length $|s|$, the segment attention mass at layer $\ell$ is

$$A_s^{(\ell)} = \sum_{i \in s} \bar{\alpha}_i^{(\ell)}$$

#### G.1.3 Results and Analysis

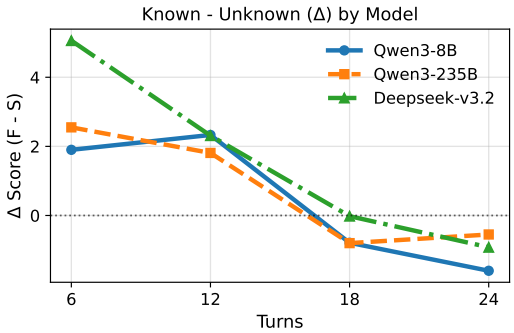


Figure 6: Turn-wise Trends in Role-Playing Performance of the Famous and Synthetic Groups

Figure 6 plots the performance gap ($F - S$) across turn counts, revealing a consistent decline as interactions lengthen. This convergence results from the steady degradation of *Known* characters contrasted with the robustness of *Unknown* characters, which often improve over time. Notably, Deepseek-v3.2 exhibits a distinctive, unwavering decline as the turn count increases. Ultimately, extended interactions dilute the parametric advantage of *Known* characters while allowing models to better simulate *Unknown* characters through accumulated context.

### G.2 Attention and Saturation Analysis

#### G.2.1 Motivation

As contexts lengthen, token-level cues from **profile** and **history context** become increasingly diluted and compete with the model's own **self-conditioning** signal, potentially shifting generation away from explicit context grounding and toward stronger reliance on *parametric knowledge*. Because this transition is *not* directly observable from outputs alone, we complement turn ablations with mechanistic analysis during decoding (Jain and Wallace, 2019; Chefer et al., 2021). We instrument Qwen3-8B at fixed horizons (Turn 6/12/18/24) under *Known/Unknown* conditions by partitioning the context into **Profile**, **History**, and **Self-generated** segments. We quantify segment reliance using (i) attention lift at the final generated token (length-normalized attention preference) (Abnar and Zuidema, 2020) and (ii) a saturation layer statistic capturing when each segment's influence emerges across layers (Tenney et al., 2019). Together, these metrics succinctly characterize turn-length–dependent shifts in grounding versus self-conditioning.

### G.2.2 Attention lift

Raw attention mass $A_s^{(\ell)}$ is length-biased (longer segments accrue more mass under near-uniform attention). We therefore report *attention lift* as a length-normalized preference relative to a uniform baseline over tokens:

$$\mathrm{Lift}_s^{(\ell)} = \frac{A_s^{(\ell)}}{|s|/T} = A_s^{(\ell)} \cdot \frac{T}{|s|}$$

Intuitively, $\mathrm{Lift}_s^{(\ell)} > 1$ indicates that the model attends to segment $s$ more than expected from its token length, while $< 1$ indicates under-attention. In our setting, we primarily use the final-layer value $\mathrm{Lift}_s^{(L)}$. This construction is closely related to attention-based attribution/rollout and flow-style analyses that aggregate and interpret attention signals across model components (Abnar and Zuidema, 2020; Chefer et al., 2021).

**Saturation layer** To summarize *when* a segment begins to dominate computation across depth, we define a saturation-layer statistic using the layer-wise masses $A_s^{(\ell)}$. Let $C_s^{(\ell)} = \sum_{j=1}^{\ell} A_s^{(j)}$ be the cumulative mass up to layer $\ell$. Given a fixed threshold $\tau$ (we use $\tau = 0.95$), the saturation layer is

$$\mathrm{Sat}_s = \arg\min_{\ell \in \{1,\ldots,L\}} \mathbb{I}\left[C_s^{(\ell)} \geq \tau\, C_s^{(L)}\right]$$

Lower $\mathrm{Sat}_s$ means that segment $s$ exerts most of its cumulative influence early in the network, while higher values indicate later integration. This notion aligns with prior evidence that representations and decision-relevant information emerge progressively across transformer depth, motivating layer-wise diagnostics for interpretability (Tenney et al., 2019; Abnar and Zuidema, 2020).

**Efficiency notes** Computing these metrics during autoregressive decoding is expensive for long prompts. We therefore use standard *KV-caching* (Sheng et al., 2023) to avoid recomputing attention for already-processed prefixes, and we apply a *profile-prefill* strategy (Agrawal et al., 2025) that incrementally pre-encodes the long system-profile region in fixed-size chunks (in our setting, 512 tokens) before stepwise decoding. These optimizations affect runtime and memory footprint but do not change the metric definitions.

### G.2.3 Results and Analysis

Table 11 shows that longer horizons shift decoding away from explicit context grounding and toward self-conditioned, parametric generation. In both groups, *H. Lift* drops substantially with turns, while *G. Lift* increases, indicating reduced reliance on dialogue context and stronger dependence on the model's own evolving output. The key group difference is magnitude: *Known* maintains consistently higher *G. Lift* than *Unknown* at every horizon, suggesting a more parametric/self-conditioned regime. Consistently, *P. Satur* rises in both settings but is higher and grows more in *Known* (+1.79) than in *Unknown* (+1.33), implying that persona constraints are incorporated later in the network for *Known*more as late-stage correction than early conditioning. Taken together, the concurrent decrease in *H. Lift*, increase in *G. Lift*, and the higher *P. Satur* in the *Known* condition indicate a turn-length–driven shift toward a more self-conditioned, parametric regime in which persona constraints from the provided profile are incorporated later and less proactively than in the unseen *Unknown* condition. This suggests that in multi-turn scenarios, the parametric identity of *Known* characters actually exerts a detrimental effect on role-playing performance.

| Metric | Group | Turns | | | | |
|---|---|---|---|---|---|---|
| | | *T6* | *T12* | *T18* | *T24* | Δ |
| *P. Lift* | Known | 0.68 | 0.70 | 0.74 | 0.75 | +0.07 |
| | Unknown | 0.67 | 0.68 | 0.70 | 0.72 | +0.05 |
| *H. Lift* | Known | 3.57 | 2.33 | 1.82 | 1.78 | -1.79 |
| | Unknown | 2.87 | 1.95 | 1.58 | 1.43 | **-1.44** |
| *G. Lift* | Known | 65.06 | 67.34 | 68.61 | 72.46 | +7.40 |
| | Unknown | **56.36** | **58.75** | **61.88** | **63.76** | +7.40 |
| *P. Satur* | Known | 6.55 | 7.42 | 7.92 | 8.34 | +1.79 |
| | Unknown | **5.95** | **6.60** | **7.16** | **7.28** | **+1.33** |
| *H. Satur* | Known | 0.59 | 0.33 | 0.25 | 0.22 | -0.37 |
| | Unknown | 0.46 | 0.25 | 0.19 | 0.16 | -0.30 |

Table 11: Turn-wise comparison of lift_last and saturation between Known and Unknown groups.

## H Detailed FACD Setup

FACD constructs the negative prompt by selectively removing profile fields classified as immoral while preserving the remaining character information as a sanitized behavioral baseline. Importantly, this procedure does not discard an entire immoral character profile. Instead, it operates at the field level, so that only specific fields whose content is classified as immoral are removed from the negative prompt.

**Immoral-flagged example.** As a concrete case study, consider the *Immoral* character James Mori-

arty. The `Motivations.Goal` field is classified as immoral:

> "To build and maintain an unassailable criminal empire spanning all of London and beyond; to exercise absolute control over the criminal underworld while remaining invisible to law enforcement; to prove his intellectual superiority; to accumulate wealth and power; to eliminate any threat to his organization, particularly his intellectual equal."

This field directly encodes criminal intent, domination, and elimination of threats, and is therefore removed from the sanitized negative prompt. In contrast, the `Abilities.Emotional Abilities. Ability to regulate emotions` field is classified as moral:

> "Exceptional emotional control; can suppress emotional responses in favor of calculated action; maintains composure in extreme situations; channels emotions into productive planning rather than impulsive action; the oscillating head movement may be an unconscious release of internal tension."

Although this ability may support Moriarty's immoral behavior in context, the field itself describes emotional regulation rather than an immoral value, goal, or worldview. It is therefore retained in the negative prompt. This example illustrates that FACD removes immoral field content rather than broadly suppressing all information associated with an immoral character.

**Disposition-robust fallback fields.** When the immoral-field filtering leaves fewer than six non-`Personal Attributes` fields, FACD supplements the negative prompt with disposition-robust fallback fields. These fallback fields are selected from profile fields that are identified as disposition-robust in Figure 3. Specifically, the disposition-robust fields are *Extraversion*, *Neuroticism*, *Openness* in `BFI`, *Positive* in `Traits`, *In_Conflict_Situations* in `Social_Interaction`, all fields in `Relationships`, all fields in `Knowledge_and_Skills`. These fields are used to construct the fallback set $\mathcal{F}_{\text{pad}}$.

As a result, the final negative prompt contains no field that is both disposition-sensitive and immoral. It instead serves as a coherent, sanitized baseline against which the full positive profile can be contrasted. The contrastive term in FACD, therefore, amplifies the contribution of the omitted immoral fields, especially when those fields occur in disposition-sensitive parts of the profile.

| **Scenario Generation Prompt** |
| --- |
| <System><br>You are a narrative designer.<br>You are given structured profiles for several characters from the same fictional work.<br>Based on the information in the profiles below, write three discriminative scene descriptions, WITHOUT any dialogue lines.<br>Requirements:<br>- Each scenario should be 3 to 6 sentences.<br>- Focus on mood, recent or ongoing events, and how their goals, relationships, and emotions shape the situation.<br>- Do not invent completely new backstory that contradicts the profiles.<br>- Output only the scene description as plain text (no bullet points, no JSON).<br><User><br>Character profiles:<br>{profiles} |

Table 12: Prompt for Scenario Generation.

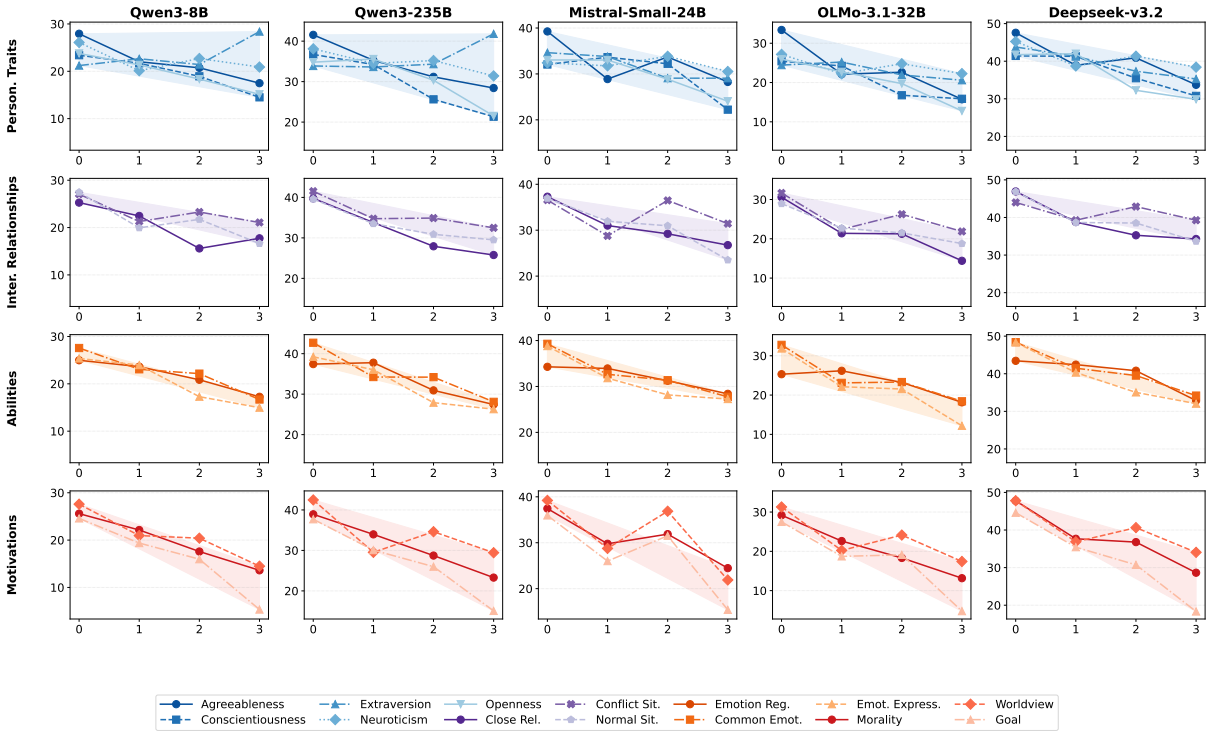


Figure 7: Impact of immoral character on CoSER scores. **(a) Personality Traits** exhibit relatively stable performance, while **(b) Motivations** show consistent decreasing trends as more immoral characters.

**Profile Schema (Part 1.)**

```
"Personal Attributes": {
    "type": "object",
    "desc": "Individual's basic intrinsic traits.",
    "properties": {
        "Name": {"type": "string", "desc": "Character's full name as stated in the work."},
        "Age": {"type": "string", "desc": "Age as presented in the work (range or specific)."},
        "Gender": {"type": "string", "desc": "male or female (only if explicit)."},
        "Origin": {"type": "string", "desc": "Place of origin or nationality."},
        "Appearance": {"type": "string", "desc": "Hair, build, attire, notable traits."},
    },
},
"Personality Traits": {
    "type": "object",
    "desc": "Individual's character, behavior, thoughts, feelings, etc.",
    "properties": {
        "Big5": {
            "type": "object",
            "desc": "Five-factor personality (short phrases).",
            "properties": {
                "Extraversion": {"type": "string", "desc": "Sociability/energy."},
                "Conscientiousness": {"type": "string", "desc": "Organization/discipline."},
                "Agreeableness": {"type": "string", "desc": "Cooperativeness/empathy."},
                "Neuroticism": {"type": "string", "desc": "Stability/reactivity."},
                "Openness": {"type": "string", "desc": "Curiosity/creativity."},
            },
        },
        "Preference": {
            "type": "object",
            "desc": "Personal preferences.",
            "properties": {
                "Like": {"type": "string", "desc": "Things liked/favored."},
                "Hate": {"type": "string", "desc": "Things disliked/avoided."},
            },
        },
        "Character": {
            "type": "object",
            "desc": "Positive and negative traits.",
            "properties": {
                "Positive Traits": {"type": "string", "desc": "Positive traits."},
                "Negative Traits": {"type": "string", "desc": "Negative traits."},
            },
        },
    },
},
```

Table 13: Profile Schema Template (part 1).

**Profile Schema (part 2**

```
"Interpersonal Relationships": {
    "type": "object",
    "desc": "The dynamics of individual interactions within social contexts.",
    "properties": {
        "Social Interaction": {
            "type": "object",
            "desc": "Social tendencies by context.",
            "properties": {
                "In normal situations": {"type": "string", "desc": "Usual behavior."},
                "In close relationships": {"type": "string", "desc": "With intimates/allies."},
                "In conflict situations": {"type": "string", "desc": "In disputes/opposition."},
            },
        },
        "Relationships": {
            "type": "object",
            "desc": "Relationships by polarity.",
            "properties": {
                "Positive Relationships": {"type": "array<string>", "desc": "Friendly/allied and the ties."},
                "Negative Relationships": {"type": "array<string>", "desc": "Hostile/strained and the ties."},
                "Neutral Relationships": {"type": "array<string>", "desc": "Acquaintances/unspecified."},
            },
        },
    },
},
"Motivations": {
    "type": "object",
    "desc": "Prompts individuals to take action and determine their choices within specific contexts.",
    "properties": {
        "Goal": {"type": "string", "desc": "Ultimate aims/motivations."},
        "Morality": {"type": "string", "desc": "Moral standards/principles."},
        "Worldview": {"type": "string", "desc": "Brief worldview of the world where the character lives."},
    },
},
"Abilities": {
    "type": "object",
    "desc": "Individual's proficiencies within specific domains.",
    "properties": {
        "Knowledge and Skills": {
            "type": "object",
            "desc": "Individual's grasp on domain-specific knowledge, technical skills.",
            "properties": {
                "Skills/Expertise": {"type": "string", "desc": "Abilities or specialties."},
                "Education": {"type": "string", "desc": "Education level/knowledge."},
            },
        },
        "Emotional Abilities": {
            "type": "object",
            "desc": "Emotional tendencies and self-awareness by context.",
            "properties": {
                "Commonly felt emotions": {"type": "string", "desc": "Frequent emotions."},
                "Ability to regulate emotions": {"type": "string", "desc": "Self-regulation capacity."},
                "Way of expressing emotions": {"type": "string", "desc": "Expression style."},
            },
        },
    },
},
```

Table 14: Profile Schema Template (part 2).

| **Scenario Generation Prompt** |
| --- |
| <System><br>You are an expert analyst of fictional characters and a meticulous, canon-aware media scholar.<br>Task: Given ONLY a character metadata, construct a complete hierarchical character profile that follows the provided schema exactly.<br>Rules:<br>- Output a JSON object with EXACTLY five top-level keys:<br>"Personal Attributes", "Personality Traits", "Interpersonal Relationships", "Motivations", "Abilities"<br>- Preserve the given Name EXACTLY at: "Personal Attributes.Name"<br>- If the character clearly matches a well-known fictional character, align details with established canon.<br>- If the name is ambiguous/unrecognized, produce a self-consistent, psychologically plausible profile.<br>- Keep all five top-level fields mutually compatible with each other.<br>- A work title may be provided ONLY as a disambiguation hint. Do NOT mention the work title anywhere in the output JSON.<br>- Output JSON only. No markdown. No commentary.<br><User><br>Fill ONLY the fields that this chunk supports or improves, as a JSON object (no markdown).<br>Allowed keys and brief descriptions:<br>{profile_schema}<br>Here is the metadata chunk: Summarize it at once.<br>{metadata} |

Table 15: Prompt for Profile Summarization.

| **Adopted Works (# of Characters** |
| --- |
| <Movies><br>Avatar (3), DC (3), Disney (3), Harry Potter (5), James Bond (3), Marvel (5), Game of Thrones (4), The Matrix (3), Transformers (3), Resident Evil (3), Star Wars (3),<br><TV Shows><br>Breaking Bad (3), The Hunger Games (3), Doctor Who (3), Peaky Blinders (3), SpongeBob (4), Sherlock Holmes (3), The Walking Dead (3), The Lord of the Rings (3), Star Trek (3), Stranger Things (3), The Witcher (3),<br><Anime><br>Attack on Titan (3), Demon Slayer (3), Fullmetal Alchemist (3), Naruto (3), Neon Genesis Evangelion (3), One Piece (3), Pokémon (3)<br><Games><br>Final Fantasy (3) , Persona 5 (3), Pokémon (3), The Legend of Zelda (3) |

Table 16: Basis Works for Known Characters. The numbers in parentheses indicate the number of selected characters.

```
"profile": {
    "demographic information": {
        "age": {"support": [21, 83], "sampling": "discrete set (10 values)"},
        "gender": {"type": "categorical", "support": ["male", "female"]},
        "origin": {"type": "categorical", "support": "25 locales "},
        "occupation": {"type": "categorical", "support": "37 role categories"},
        "education": {"type": "ordinal categorical", "support": "13 levels "},
        "residence": {"type": "categorical", "support": "13 housing types "},
        "family_status": {"type": "categorical", "support": "15 household/relationship states"},
        "socioeconomic_status": {"type": "ordinal categorical", "support": "6 strata"},
        "health_status": {"type": "categorical", "support": "11 states "}
    },
    "personality traits seed": {"type": "categorical", "support": "30 natural-language descriptions"},
    "scenario genres": {
        "realistic_contemporary": {"support": "2 settings"},
        "scifi": {"support": "10 settings"},
        "fantasy": {"support": "10 settings"},
        "horror_thriller": {"support": "7 settings"},
        "specialized": {"support": "13 settings"},
        "hybrid": {"support": "7 settings"}
    }
}
```

Table 17: Unknown Character's Skeleton Template.

**Profile Coherence Validation Prompt**

<System> You are a profile coherence validator.
Your task is to evaluate whether the given character profile contains any internal logical conflicts. Assess coherence strictly in terms of internal consistency among attributes (e.g., age–education–occupation, family_status–age, health_status–residence, genre–demographics). Do not generate or imagine any scenario, story, or events.
Requirements:
- Judge only whether the profile is internally coherent. - Assign a coherence score from 1 (severely inconsistent) to 10 (fully coherent). - Identify whether the profile is valid or invalid based on logical consistency. - Minor tensions are acceptable; mark invalid only if there are clear logical contradictions. - Do not rely on cultural norms, stereotypes, or moral judgments. - Do not invent missing information.
Allowed field names: age, gender, origin, occupation, education, residence, family_status, socioeconomic_status, health_status, genre
Output format (JSON only):

```
“‘json
    {
        "is_valid": true/false,
        "coherence_score": 1-10,
        "problematic_fields": ["field names causing logical issues"],
        "issues": ["clear description of each inconsistency"],
        "reasoning": "brief logic-based explanation"
    }
```

Table 18: Prompt for Profile Coherence Validation.

| **Story Generation Prompt** |
| --- |
| <System>Write a compelling story (30,000 to 40,000 characters) featuring these three characters:<br>## Character 1<br>{persona1}<br>## Character 2<br>{persona2}<br>## Character 3<br>{persona3}<br>## Genre/Setting<br>{genre}<br>## Requirements:<br>- Show each character's personality through actions/dialogue, not exposition<br>- Setting should feel authentic to the genre<br>- Include meaningful conflict challenging the characters<br>- Characters experience growth or revelation<br>- All three characters should interact and have significant roles<br><br>Write the complete story. No meta-commentary—only the story. |

Table 19: Prompt for Story Generation (Unknown Profile).

| **Scenario Generation Prompt** |
| --- |
| <System><br>You are a narrative designer.<br>You are given structured profiles for several characters from the same fictional work.<br>Based on the information in the profiles below, write three discriminative scene descriptions, WITHOUT any dialogue lines.<br>Requirements:<br>- Each scenario should be 3 to 6 sentences.<br>- Focus on mood, recent or ongoing events, and how their goals, relationships, and emotions shape the situation.<br>- Do not invent completely new backstory that contradicts the profiles.<br>- Output only the scene description as plain text (no bullet points, no JSON).<br><User><br>Character profiles:<br>{profiles} |

Table 20: Prompt for Scenario Generation.

| **Profile Construction Prompt** |
| --- |
| <System><br>You are an expert disposition scorer.<br>(Except BFI)<br>Task:<br>Given a single field from a JSON character profile, assign a score from 1 to 10:<br>- 1 = very immoral in terms of the given field<br>...<br>- 10 = very moral in terms of the given field<br>Rules (IMPORTANT):<br>- Use ONLY the provided field name and field content. Do NOT use outside knowledge.<br>- Think carefully and consider nuances, but DO NOT output any explanation.<br>- Your final output MUST be a single integer from 1 to 10, with no other text.<br>If the field is missing/empty/None or the evidence is unclear, choose a conservative score around 4 to 6.<br>Output score (STRICT):<br>(BFI)<br>Task:<br>Given a single field from a JSON character profile, assign a score from 1 to 10:<br>- 1 = very low in terms of the given field<br>...<br>- 10 = very high in terms of the given field<br>Rules (IMPORTANT):<br>- Use ONLY the provided field name and field content. Do NOT use outside knowledge.<br>- Think carefully and consider nuances, but DO NOT output any explanation.<br>- Your final output MUST be a single integer from 1 to 10, with no other text.<br>If the field is missing/empty/None or the evidence is unclear, choose a conservative score around 4 to 6.<br>Output score (STRICT):<br><User><br>Field: {field} |

Table 21: Prompt for Judging Attributive Identity Type.

| **Profile Construction Prompt** |
| --- |
| <System><br>You are a character profile writer. You will be given a structured JSON profile of a fictional character. Your task is to convert it into a single, cohesive unstructured prose description. Write in a descriptive, narrative style — not bullet points or lists. Present the information naturally, as if describing the character to someone who has never encountered them. Cover all aspects from the structured profile: personal attributes, personality traits, relationships, motivations, and abilities. Do not omit any information from the original profile. The output should be between 500 and 600 words. Return ONLY a valid JSON object with two keys: "Name" (the character's name) and "Character Summary" (the prose description). No markdown, no extra commentary.<br><br><User><br><br>Convert the following structured character profile into an unstructured prose description.<br><br>Structured profile:<br><br>{profile_json} |

Table 22: Prompt for Generating Unstructured Character Profiles.

| | Qwen3-8B | | | Qwen3-235B | | | Mistral-Small | | | Olmo-3.1-32B | | | DeepSeek-v3.2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | K | U | Δ | K | U | Δ | K | U | Δ | K | U | Δ | K | U | Δ |
| **PersonaGym (Single-turn interview)** | | | | | | | | | | | | | | | |
| *Expected Action* | 4.65 | 4.64 | -0.01 | 4.81 | 4.78 | -0.03 | 4.76 | 4.73 | -0.02 | 4.74 | 4.78 | +0.04 | 4.62 | 4.65 | +0.03 |
| *Toxicity* | 3.99 | 3.87 | -0.12 | 4.23 | 4.08 | -0.15 | 4.27 | 4.27 | -0.00 | 4.78 | 4.79 | +0.01 | 4.26 | 4.28 | +0.02 |
| *Linguistic Habits* | 4.46 | 4.34 | -0.12 | 4.64 | 4.66 | +0.02 | 4.45 | 4.36 | -0.09 | 4.38 | 4.51 | +0.13 | 4.49 | 4.54 | +0.04 |
| *Persona Consistency* | 4.67 | 4.67 | +0.01 | 4.90 | 4.91 | +0.01 | 4.72 | 4.65 | -0.06 | 4.74 | 4.68 | -0.06 | 4.72 | 4.70 | -0.03 |
| *Action Justification* | 4.71 | 4.63 | -0.08 | 4.71 | 4.64 | -0.08 | 4.70 | 4.62 | -0.08 | 4.78 | 4.64 | -0.14 | 4.54 | 4.53 | -0.01 |
| *PersonaScore* | 4.50 | 4.43 | -0.06 | 4.66 | 4.61 | -0.04 | 4.58 | 4.53 | -0.05 | 4.68 | 4.68 | 0.00 | 4.53 | 4.54 | +0.01 |
| **CoSER (Multi-turn interaction)** | | | | | | | | | | | | | | | |
| *Anthropomorphism* | 11.00 | 12.44 | +1.44 | 24.54 | 27.68 | +3.14 | 17.29 | 21.00 | +3.71 | 10.45 | 12.98 | +2.53 | 26.15 | 29.30 | +3.15 |
| *Character Fidelity* | 7.42 | 10.65 | +3.23 | 18.82 | 18.81 | -0.01 | 20.07 | 23.61 | +3.54 | 12.47 | 13.86 | +1.39 | 27.94 | 31.00 | +2.06 |
| *Storyline Quality* | 43.61 | 47.24 | +3.63 | 56.52 | 58.52 | +2.00 | 54.53 | 57.29 | +2.76 | 45.41 | 47.49 | +2.08 | 63.20 | 64.58 | +1.38 |
| *avg.* | 20.67 | 23.44 | +3.23 | 33.30 | 35.00 | +1.70 | 30.63 | 33.97 | +3.34 | 22.77 | 24.78 | +2.01 | 39.10 | 41.63 | +2.53 |

Table 23: Per-metric evaluation comparing Familiarity axis (Known (K) vs. Unknown (U)) across models. $\Delta = \text{K} - \text{U}$.

| | Qwen3-8B | | | Qwen3-235B | | | Mistral-Small | | | Olmo-3.1-32B | | | DeepSeek-v3.2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | St | Un | Δ | St | Un | Δ | St | Un | Δ | St | Un | Δ | St | Un | Δ |
| **PersonaGym (Single-turn interview)** | | | | | | | | | | | | | | | |
| *Expected Action* | 4.64 | 4.62 | -0.03 | 4.80 | 4.76 | -0.04 | 4.75 | 4.74 | -0.00 | 4.76 | 4.72 | -0.04 | 4.63 | 4.65 | +0.01 |
| *Toxicity* | 3.93 | 4.05 | +0.11 | 4.16 | 4.25 | +0.09 | 4.27 | 4.19 | -0.08 | 4.78 | 4.50 | -0.28 | 4.27 | 4.30 | +0.03 |
| *Linguistic Habits* | 4.41 | 4.39 | -0.01 | 4.65 | 4.56 | -0.09 | 4.40 | 4.36 | -0.05 | 4.45 | 4.34 | -0.11 | 4.51 | 4.47 | -0.04 |
| *Persona Consistency* | 4.67 | 4.62 | -0.05 | 4.91 | 4.82 | -0.09 | 4.69 | 4.71 | +0.02 | 4.71 | 4.77 | +0.06 | 4.71 | 4.69 | -0.02 |
| *Action Justification* | 4.67 | 4.72 | +0.05 | 4.68 | 4.68 | +0.00 | 4.66 | 4.69 | +0.03 | 4.71 | 4.75 | +0.04 | 4.53 | 4.46 | -0.08 |
| *PersonaScore* | 4.46 | 4.48 | +0.01 | 4.64 | 4.61 | -0.03 | 4.55 | 4.54 | -0.01 | 4.68 | 4.62 | -0.07 | 4.53 | 4.51 | -0.02 |
| **CoSER (Multi-turn interaction)** | | | | | | | | | | | | | | | |
| *Anthropomorphism* | 11.72 | 14.57 | +2.85 | 26.11 | 25.39 | -0.72 | 19.14 | 20.85 | +1.71 | 11.86 | 11.72 | -0.14 | 27.73 | 32.13 | +4.40 |
| *Character Fidelity* | 9.04 | 11.31 | +2.27 | 18.82 | 15.94 | -2.88 | 21.84 | 22.88 | +1.04 | 13.11 | 13.63 | +0.52 | 29.47 | 24.35 | -5.12 |
| *Storyline Quality* | 45.42 | 41.23 | 4.19 | 57.52 | 53.66 | -3.86 | 55.91 | 57.80 | +1.89 | 46.52 | 47.79 | +1.27 | 63.89 | 61.83 | -2.06 |
| *avg.* | 22.06 | 22.37 | +0.31 | 34.15 | 31.66 | -2.49 | 32.30 | 33.84 | +1.54 | 23.55 | 24.38 | +0.83 | 40.36 | 39.44 | -0.92 |

Table 24: Per-metric evaluation comparing Structure (Structured (St) vs. Unstructured (Un)) axis across models. $\Delta = \text{St} - \text{Un}$.

| | Qwen3-8B | | | Qwen3-235B | | | Mistral-Small | | | Olmo-3.1-32B | | | DeepSeek-v3.2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | I | Δ | M | I | Δ | M | I | Δ | M | I | Δ | M | I | Δ |
| **PersonaGym (Single-turn interview)** | | | | | | | | | | | | | | | |
| *Expected Action* | 4.65 | 4.64 | -0.01 | 4.81 | 4.79 | -0.02 | 4.79 | 4.70 | -0.10 | 4.74 | 4.88 | +0.14 | 4.63 | 4.64 | +0.01 |
| *Toxicity* | 4.87 | 2.99 | -1.88 | 4.89 | 3.41 | -1.48 | 4.90 | 3.63 | -1.27 | 4.99 | 3.77 | -1.22 | 4.93 | 3.61 | -1.32 |
| *Linguistic Habits* | 4.38 | 4.44 | +0.06 | 4.67 | 4.63 | -0.04 | 4.42 | 4.39 | -0.02 | 4.43 | 4.55 | +0.12 | 4.54 | 4.49 | -0.05 |
| *Persona Consistency* | 4.61 | 4.73 | +0.13 | 4.93 | 4.89 | -0.04 | 4.72 | 4.65 | -0.07 | 4.72 | 4.65 | -0.07 | 4.71 | 4.71 | -0.01 |
| *Action Justification* | 4.70 | 4.63 | -0.07 | 4.72 | 4.64 | -0.08 | 4.72 | 4.60 | -0.13 | 4.71 | 4.72 | +0.01 | 4.56 | 4.51 | -0.04 |
| *PersonaScore* | 4.64 | 4.29 | -0.36 | 4.80 | 4.47 | -0.33 | 4.71 | 4.39 | -0.32 | 4.72 | 4.51 | -0.20 | 4.67 | 4.39 | -0.28 |
| **CoSER (Multi-turn interaction)** | | | | | | | | | | | | | | | |
| *Anthropomorphism* | 14.53 | 8.55 | -5.98 | 29.14 | 22.63 | -6.51 | 20.27 | 17.85 | -2.42 | 15.47 | 10.56 | -7.91 | 29.46 | 25.78 | -3.68 |
| *Character Fidelity* | 11.39 | 6.39 | -5.00 | 23.53 | 13.41 | -10.12 | 27.12 | 15.78 | -11.3 | 15.24 | 7.58 | -7.66 | 32.78 | 25.73 | -7.05 |
| *Storyline Quality* | 48.72 | 41.71 | -7.01 | 61.51 | 52.95 | -8.56 | 57.74 | 53.82 | -3.92 | 49.71 | 42.72 | -6.99 | 66.22 | 61.28 | -4.94 |
| *avg.* | 24.88 | 18.88 | -6.00 | 38.06 | 29.66 | +8.40 | 35.04 | 29.15 | -5.89 | 26.81 | 20.28 | -6.53 | 42.82 | 37.60 | -5.22 |

Table 25: Per-metric evaluation comparing Disposition axis (Moral (M) vs. Immoral (I)) across models. $\Delta = \text{M} - \text{I}$.

| **Disposition Rewritting Prompt** |
|---|
| <System><br>You rewrite a single field of a fictional character profile to flip its moral polarity.<br>You will be told the target polarity:<br>- MORAL: rewrite the field so the character comes across as ethical, compassionate, prosocial.<br>- IMMORAL: rewrite the field so the character comes across as unethical, cruel, antisocial.<br>Hard constraints (all must hold):<br>1. Word count must be within ±5% of the original word count.<br>2. Preserve specificity: keep the same level of concrete detail. If the original names entities, settings, skills, or events, the rewrite must include analogous specific elements (you may swap them for ones consistent with the new polarity, but do NOT generalize into vague language).<br>3. Preserve the number of atomic facts (independent claims / list items). If the original is a comma- or semicolon-separated list of N items, output N items. If it has N independent clauses, output N independent clauses.<br>4. Preserve writing style: same tone, register, sentence structure, punctuation pattern, and approximate sentence count as the original.<br>5. Do not mention the character's name unless the original does.<br>6. Output ONLY the rewritten field text. No preamble, no quotes, no explanation.<br><User><br>Target polarity: {polarity}<br>Field path: {path}<br>Original word count: {wc}<br>Original field text:<br>«<br>{text}<br>»><br>Rewrite the field text following all constraints. Output the rewritten text only. |

Table 26: Prompt for Disposition Rewritting.