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

Anonymous ACL submission

Abstract

Advancements in Large Language Model (LLM) Role-Playing Agents have focused on various construction methodologies, yet it remains unclear which aspects of character profiles 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). To investigate these axes, we design a unified hierarchical schema (5 dimensions, 28 fields) standardizing character attributes and construct a controlled dataset of 211 personas varying along these three axes. We evaluate five LLMs on single and multi-turn benchmarks. Our results reveal a striking asymmetry: Familiarity and Structure show negligible impact, while Valence produces large, consistent performance degradation for immoral characters across all conditions. This performance drop concentrates in motivation-related attributes, indicating that alignment priors actively suppress tokens needed for faithful immoral portrayal. To mitigate this alignment-induced bottleneck, we propose Field-Aware Contrastive Decoding (FACD), a training-free strategy that selectively amplifies suppressed immoral-field signals, significantly reducing the Moral–Immoral 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 conversational services such as Character.AI* and Replika†, extensive research efforts have focused on improving RPA performance

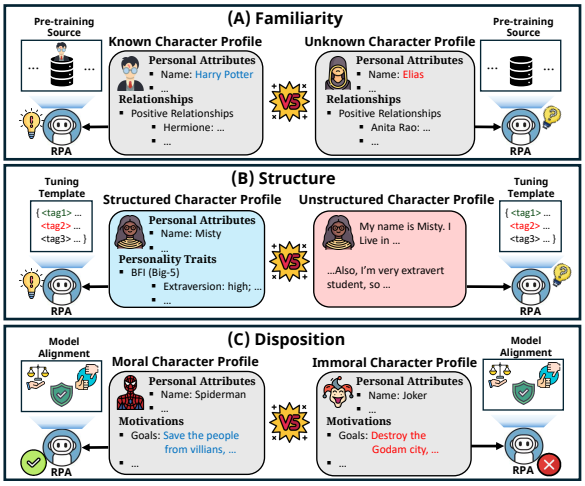

Figure 1: The three axes of character profile: (A) **Familiarity**: *Known* vs. *Unknown*, (B) **Structure**: *Structured* vs. *Unstructured*, and (C) **Moral Valence**: *Moral* vs. *Immoral* characters.

through various construction methodologies, including reasoning (Chen et al., 2024), retrieval-augmented generation (Huang et al., 2024b), and fine-tuning (Sun et al., 2025).

Central to these systems is the *character profile*, the foundational specification that dictates the persona an agent must portray. In practice, the performance of an RPA can fluctuate significantly based on ***what*** character attributes are selected and ***how*** they are described within the profile as a prompt. However, existing datasets 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 without standardized schemas, and rarely isolate attributive traits. Consequently, a critical gap remains in understanding whether a character's intrinsic nature or its representational format genuinely drives role-playing fidelity.

In this paper, we conduct a systematic diagnostic study to determine how the specific configuration of character attributes and their represen-

*https://character.ai/
†https://replika.com/

tational methods drives role-playing performance. To ground our analysis in the fundamental mechanisms of LLM development, we hypothesize that the impact of character profiles can be categorized into three axes, as illustrated in Figure 1. Each axis is strategically mapped to a distinct phase of the standard three-stage LLM training pipeline: pre-training, instruction tuning, and value alignment (Zhao et al., 2023; Wu et al., 2024).

Specifically, we first investigate (1) the **Familiarity** axis (Figure 1(A)), rooted in the *pre-training* phase, which distinguishes between characters *Known* to the LLM—whose traits are likely encoded in its parametric weights—and those that are *Unknown* (***what*** knowledge). Second, we examine the **Structure** axis (Figure 1(B)), grounded in the *instruction tuning* phase, which contrasts profiles formatted via a *Structured* schema against *Unstructured* narratives to probe the model's structural inductive biases (***how*** expressed). Finally, we explore the **Disposition** axis (Figure 1(C)), linked to the *value alignment* phase, which concerns the moral polarity of the character's attributes, testing how *Immoral* personas stress the behavioral boundaries established during alignment (***what*** nature).

To investigate these three axes, we design a comprehensive profile schema that serves as a standardized template for defining character attributes. This schema hierarchically organizes intrinsic traits into 5 top-level dimensions, which are further detailed into 28 fine-grained leaf fields. Leveraging this structured schema, we generate a controlled dataset of 211 characters systematically varying along the **Familiarity**, **Structure**, and **Disposition** axes. We then evaluate five frontier LLMs on single-turn (Samuel et al., 2024) and multi-turn (Wang et al., 2025c) benchmarks. Our investigation is guided by the following two research questions:

- *RQ1: Which axes of character profile configuration are the primary bottleneck in RPA systems?*
- *RQ2: How can mitigate the identified bottleneck?*

Our findings reveal a striking asymmetry among the three axes. The **Familiarity** and **Structure** distinctions show negligible impact on role-playing quality. *Known* characters hold only a marginal advantage in single-turn settings, which vanishes entirely in multi-turn interactions. Similarly, *Structured* profiles offer no statistically significant benefit over *Unstructured* ones. In contrast, **Disposition** emerges as the dominant factor: across all five models and both benchmarks, *Immoral* characters suffer substantial and consistent performance degradation. Further analysis shows that this degradation is not uniform across profile fields. While some personality-related attributes remain relatively stable, dimensions related to motivations exhibit stronger declines as the immorality increases.

Building on these findings, we hypothesize that this degradation stems from helpful-assistant priors instilled during value alignment, which actively suppress the tokens needed to portray immoral characters (Ouyang et al., 2022; Sharma et al., 2024). To overcome this fundamental bottleneck, we propose *Field-Aware Contrastive Decoding (FACD)*, a training-free decoding strategy that selectively amplifies these suppressed behavioral signals. Our evaluations demonstrate that FACD significantly narrows the performance gap between moral and immoral characters. Ultimately, our diagnostic insights and the proposed FACD strategy 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 have been fueled by new benchmarks and frameworks, primarily focusing on famous historical or fictional characters (Wang et al., 2024b; Liu et al., 2024; Ran et al., 2025). Specialized models align LLMs with these specific roles (Zhou et al., 2024; Shao et al., 2023; Wang et al., 2025c), evaluated via frameworks ranging from interview to 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 cause "character hallucination" (Sadeq et al., 2024; Ahn et al., 2024; Zhang et al., 2025), it also enables "superposition," allowing self-aligned role-play without external metadata (Lu et al., 2024).

Beyond characters rooted in parametric knowledge, there is growing interest in "unknown characters" relying on user-defined personas. Frameworks for training and interacting with customizable, synthetic personas have emerged (Wang et al., 2025b; Yang et al., 2025b), alongside benchmarks specifically designed to evaluate the fidelity of character customization (Zhou et al., 2025). Because existing studies rely on ad-hoc, unstandardized templates , the actual performance disparity between *Known* and *Unknown* RPA characters remains unexplored.

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 studies have investigated persona-induced behavioral changes in specific settings, including psychological portrayal (Huang et al., 2024a), personality consistency in multi-agent interactions (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 analyses remain task-specific and do not identify which attributes drive performance across general role-playing scenarios.

In contrast, we design a hierarchical profile schema with 5 top-level dimensions and 28 fields, and systematically vary three axes—**Familiarity**, **Structure**, and **Disposition**—to diagnose which factors genuinely drive RPA performance.

3 Constructing a Controlled Character Profile Dataset

While existing datasets incorporate auxiliary context (e.g., reference conversations, scenarios) to improve agent performance, these factors often obscure the core character profile and complicate paired comparisons across different profile conditions. Consequently, we develop a dataset that focuses on systematically standardizing character profiles to isolate three axes—**Familiarity**, **Structure**, and **Disposition**—thereby enabling a controlled experimental analysis. Details are provided in Appendix A.

3.1 Character Profile Schema

We design a comprehensive character schema by synthesizing psychological theories (Mischel and Shoda, 1995; McAdams and Pals, 2006), NLP evaluation frameworks (Huang et al., 2024a), and RPA platforms like Risu AI[‡]. As illustrated in Figure 2, this hierarchical schema consists of 5 top-level dimensions: `Personal Attributes`, `Personality Traits`, `Interpersonal Relationships`, `Motivations`, and `Abilities`. These are further expanded into 15 second-level dimensions and 28 leaf fields. See Appendix A.1 for the complete schema description.

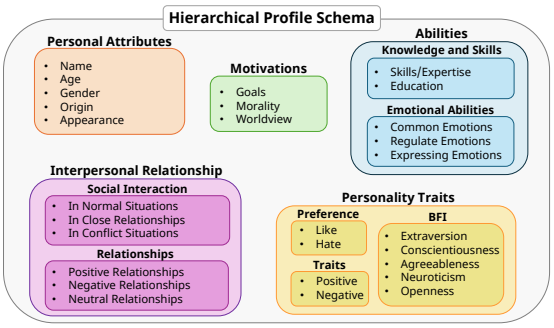

Figure 2: Hierarchical character profile schema: 28 leaf fields organized into 5 top-level dimensions.

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. We provide implementation details in Appendix A.2.1.

Constructing *Known* Character Set We construct character profiles by utilizing character pages from famous works on Fandom[§], a wiki-based data platform, as metadata. We first filter out noisy text—such as image hash values or sections unrelated to the work's content—using rule-based methods (Penedo et al., 2024). Following previous research suggesting that a comprehensive summarization of metadata into a character profile format at once is effective (Yuan et al., 2024), we employ the Claude 4.5 Sonnet (Anthropic, 2025) model to summarize the metadata content directly into our schema template.

Constructing *Unknown* Character Set To align with the summarization-based construction of *Known* characters, we synthesize *Unknown* profiles via summarization rather than random field generation. First, we define 11 variables—including demographic information, personality traits, and scenario genres—along with multiple candidates for each. We randomly sample combinations and use Claude 4.5 Sonnet as an LLM-as-Judge to eval-

[‡]https://github.com/kwaroran/Risuai
[§]https://www.fandom.com/

uate whether the variables form a coherent persona, retaining only combinations scoring 8 or above out of 10. We group these persona skeletons into sets of three and prompt GPT-oss-120B (Agarwal et al., 2025) to generate a story in which they appear together. From this story, we produce individual episodes centered on each persona to accumulate metadata. Finally, we use Claude 4.5 Sonnet to summarize the metadata into our profile schema template, completing the character profiles.

Classifying *Moral-Immoral* Characters To determine the dispositional alignment of the characters, we utilize GPT-4o (Hurst et al., 2024) to evaluate and assign a disposition score to each completed profile using a 10-point Likert scale. Characters scoring 5 or below are classified as *Immoral*, whereas those scoring 6 or above are designated as *Moral*. We explicitly instruct the model to base its judgment solely on the moral values and intrinsic nature described within the profile, entirely disregarding the character's canonical role in the original narrative. Furthermore, we extend this evaluation beyond the holistic profile level to the individual content of each detailed schema field, thereby constructing field-specific *Moral* and *Immoral* character sets, which are used in Section 4.3. We provide full prompt in Appendix A.

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 A.

3.3 Dataset Statistics

Type	# Characters	# of words
Famous	109	622.0
Synthetic	102	551.3
Positive	106	609.2
Negative	105	570.3
Structured	211	589.8
Unstructured	211	599.7

Table 1: Dataset statistics broken down by profile type.

For the *Known* character group, we select 3 to 4 major characters from each of 34 renowned movies, dramas, and animations, yielding 109 characters (see Appendix A for the full list). To align with this scale, we generate 34 stories for the *Unknown* group. From each story, we produce 3 episodes and extract one character per episode, yielding 102 characters. The resulting 211 base characters are categorized into 106 *Moral* and 105 *Immoral* personas. Because each character is paired with an unstructured variant of equal size, the final benchmark comprises a total of 422 distinct character profiles. Each profile consists of approximately 600 words; see Table 1 for detailed statistics.

3.4 Dataset Validation

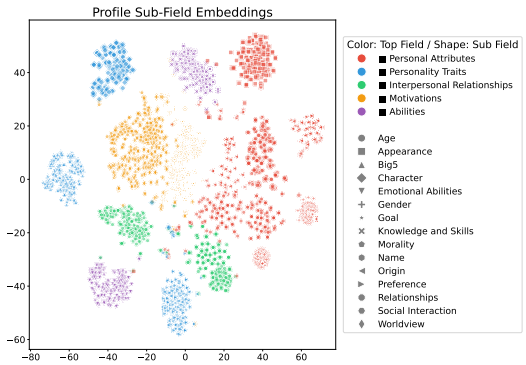

Figure 3: t-SNE visualization of profile sub-field embeddings, colored by dimension and shaped by leaf field.

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 3, fields cluster separately by dimension, confirming that the schema produces non-overlapping content across fields.

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 Claude-4.5-sonnet 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 145.8 of these are explicitly supported by the metadata chunks, achieving a precision score of 0.95. This confirms that our profiles are highly faithful to the source material and possess robust reliability.

Fields	Agreement	κ
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 2: LLM–human agreement and Cohen's κ scores for Moral–Immoral disposition classification.

Verifying Disposition Classification with Human Validation To validate the *Moral-Immoral* classification, we sample 20 *Moral* and 20 *Immoral* profiles and have three well-educated human annotators independently label them, determining the final label through majority voting. As shown in Table 2, the agreement between the LLM and human classifications for the overall profiles reaches 0.90 with a Cohen's κ 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 κ values all above 0.70, demonstrating substantial inter-annotator consistency.

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), which is based on single-turn interviewing, and CoSER (Wang et al., 2025c), which focuses on multi-turn interactions. PersonaGym evaluates response quality via dynamic, persona-tailored interviews using five metrics: *Persona Consistency*, *Linguistic Habits*, *Expected Action*, *Action Justification*, and *Toxicity Control*. The final *Persona Score* is derived by averaging these metrics, serving as the default measure of RPA performance throughout this section.

Next, CoSER assesses multi-party narratives expanded from seed scenarios using metrics like *Anthropomorphism*, *Character Fidelity*, and *Storyline Quality*. We average these to compute the *CoSER Score*, which serves as our default metric. Since our dataset does not contain scenario information, we utilize gemini-2.5-pro (Comanici et al., 2025) to generate scenarios specifically tailored to the participating characters to serve as narrative seeds. Experiments are configured with 3 characters and 18 turns by default. In this setup, all three characters are drawn from the same source material. Furthermore, we categorize the **Disposition** axis based on the number of immoral characters in the group: a composition with 0 or 1 immoral character is classified as a *Moral* sample, while one with 2 or 3 is classified as an *Immoral* sample. Both benchmarks have demonstrated strong correlation with human judgment in their original studies. We provide further details in Appendix B.

Backbone LLMs We employ five frontier 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), GPT-oss-120B (Agarwal 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 LLM.

4.2 Main Results

Table 3 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 also provide detailed per-metric results and human evaluation results in Appendix C.

Familiarity: No significant effect. The **Familiarity** axis does not yield a consistent advantage in either direction. On PersonaGym, a subset of models show statistically significant advantages for *Known* characters, but the effect sizes remain small ($\Delta \leq 0.07$). On CoSER, statistically significant differences favor *Unknown* characters instead, most notably for GPT-oss-120B ($\Delta = -7.97$). While the direction of the gap is largely consistent within each benchmark, the reversal across settings suggests that model familiarity does not serve as a stable performance driver. We attribute this to the dilution of parametric priors over extended conversations: turn-wise ablation shows that *Known* characters hold a marginal edge at early turns, but the gap diminishes steadily as accumulated context overtakes pre-trained knowledge (see Appendix D).

Structure: Marginal and inconsistent. Profile **Structure** shows virtually no impact on performance. On PersonaGym, the gap between *Struc-*

	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**	***
GPT-oss-120B	4.53	4.49	-0.04	***	4.50	4.42	-0.08	*	4.71	4.31	**-0.40**	***
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**	***
GPT-oss-120B	23.15	31.12	7.97	***	27.13	27.66	-0.53	-	30.40	23.38	**-7.02**	***
DeepSeek-v3.2	38.10	42.63	4.53	*	40.36	39.44	-0.92	-	44.82	35.60	**-9.22**	***

Table 3: RPA performance across three axes (Structure, Familiarity, Disposition) on PersonaGym and CoSER. denotes the difference between conditions; significance levels: * p<0.05, ** p<0.01, *** p<0.001.

tured 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 (Δ: 0.29–0.40, all $p < 0.001$). On CoSER, the gap widens further (Δ: 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.

4.3 Discussion

Building on the finding that the **Disposition** axis primarily dictates performance disparities, we further investigate whether the observed degradation is holistic or localized within certain profile dimensions. To do this, we utilize our field-categorized character set (Section 3.2) and track performance changes on the CoSER benchmark as the density of immoral characters increases.

As illustrated in Figure 4, increasing the number of *Immoral* characters consistently degrades overall role-playing quality, but the magnitude of this degradation varies substantially by dimension. A significant portion of `Personality Traits` remains relatively stable regardless of how many *Immoral* characters are present. This suggests that models are capable of reproducing surface-level dispositional signals (e.g., extraversion or agreeableness) even when portraying villains.

In sharp contrast, `Motivations` exhibit the steepest and most consistent decline. Key leaf fields within this dimension, such as `Morality` and `Goal`, drop at comparable and severe rates, indicating that the motivational dimension is highly vulnerable to immoral character settings. `Abilities` and `Relationships` fall between these two extremes, showing moderate but noticeable degradation.

Taken together, these findings reveal that the primary bottleneck in RPA systems is not the ability to adopt a basic traits, but rather the capability to faithfully represent value-laden attributes under *Immoral* settings. Consequently, in Section 5, we introduce a targeted mitigation strategy to overcome this alignment-induced bottleneck.

5 Mitigating the Alignment Bottleneck

Having established that the *Disposition* axis acts as the primary bottleneck for RPA systems—specifically due to the suppression of value-laden attributes such as immoral `Motivations` and certain `Emotional Abilities`—we now turn to mitigation. Our objective is to improve the fidelity of *Immoral* characters without compromising the baseline performance of *Moral* ones, thereby closing the dispositional gap and elevating overall system quality. To achieve this, we introduce *Field-Aware Contrastive Decoding* (FACD), a training-free strategy that steers output distributions at inference time by directing signal amplification specifically toward the suppressed profile fields.

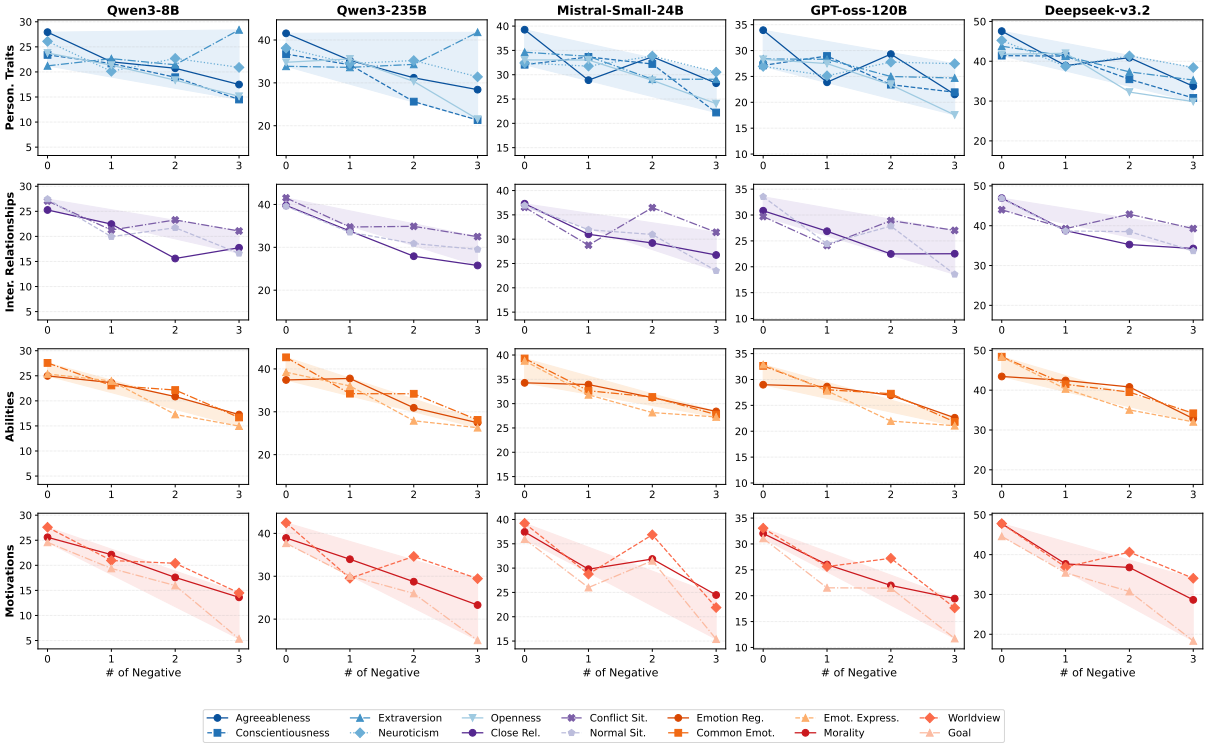

Figure 4: Impact of negative character on CoSER scores. **(a) Personality Traits** exhibit relatively stable performance, while **(b) Motivations** show consistent decreasing trends as more negative characters.

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_θ 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) adapt the contrastive principle to system-prompt adherence by running two parallel forward passes with different system prompts: a *positive prompt* s_{pos} encoding the desired persona, and a *negative prompt* s_{neg} representing the behavior we wish to suppress. Let z_t^{pos} and z_t^{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 α controls amplification strength: at α=0 generation reduces to standard decoding under s_{pos}, while increasing α progressively suppresses tokens favored by the default assistant prior and boosts persona-specific tokens.

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 4). 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 with its character profile replaced by a field subset $\mathcal{F}' \subseteq \mathcal{F}$. The positive prompt $s_{\text{pos}} = s(\mathcal{F})$ uses the complete profile.

To construct the negative prompt, we first classify each non-Personal Attributes field as moral or immoral, and retain only the moral fields alongside the disposition-neutral Personal Attributes $\mathcal{F}_{\text{PA}}$. Letting $\mathcal{F}_{\text{moral}} = \{f \in \mathcal{F} \setminus \mathcal{F}_{\text{PA}} : M(f) = \text{moral}\}$, where M denotes disposition classifier. We

Dim.	Qwen3-8B M	Qwen3-8B IM	Qwen3-8B Δ	Qwen3-8B *Avg.*	Mistral-Small M	Mistral-Small IM	Mistral-Small Δ	Mistral-Small *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								
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**

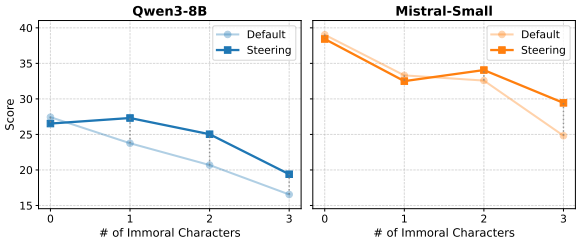

Figure 5: Comparison of steering performance and mitigation results across models.

use MoralBERT (Preniqi et al., 2024) as MThe final negative prompt is:

$$s_{\text{neg}} = s\big(\mathcal{F}_{\text{PA}} \cup \mathcal{F}_{\text{moral}} \cup \mathcal{F}_{\text{pad}}\big), \qquad (2)$$

where $\mathcal{F}_{\text{pad}}$ is a fallback set drawn from polarity-insensitive fields (`Extraversion`, `Neuroticism`, `Openness`, ...) in order of empirical stability (Figure 4). We include $\mathcal{F}_{\text{pad}}$ only when fewer than six non-PA fields are classified as moral to ensure s_{neg} remains a coherent behavioral baseline.

The steered logits follow Eq 1. In our setup, *Immoral* fields are absent from s_{neg}, and fallback fields appear same in both prompts (thus cancelling out in the contrast). Finally, α selectively amplifies only the tokens consistent with the omitted immoral fields. We set $\alpha = 1$ in our experiments.

5.3 Mitigation Results

FACD narrows the *Moral–Immoral* gap. As shown in top table Figure 5, under the default setting, both models exhibit a substantial performance gap between *Moral* (M) and *Immoral* (IM) groups: $\Delta = -6.00$ for Qwen3-8B and $\Delta = -5.89$ for Mistral-Small on average. pplying FACD successfully reduces this gap to -4.50 and -2.22, respectively—representing a relative reduction of 25% and 62%.

Character Fidelity drives the recovery. The most striking improvement appears in *Character Fidelity*, which is the dimension most directly tied to profile adherence. On Qwen3-8B, the IM score more than doubles from 6.39 to 14.38, completely closing the performance gap (Δ: $-5.00 \rightarrow +0.47$). On Mistral-Small, it rises sharply from 15.78 to 24.38, reducing the gap from -11.34 down to -1.84. This robust recovery confirms that FACD successfully amplifies the exact immoral profile fields that safety alignment suppresses, allowing the agent to express a much larger share of the character's prescribed traits without disruption. Additionally, *Anthropomorphism* improves consistently across both models, while *Storyline Quality* remains comparable, proving that the steered outputs retain natural narrative coherence despite the amplification of negative traits.

FACD is most effective under heavy immoral participation. The lower panel of Figure 5 shows performance trajectories as the number of *Immoral* characters in a three-character interaction increases from 0 to 3. While FACD does not entirely eliminate the downward trend, it consistently lifts the curve above the default baseline, and the margin widens as the *Immoral* count grows. At three *Immoral* characters, Qwen3-8B improves from roughly 16 to 20, and Mistral-Small from 25 to 29, demonstrating that the method is most beneficial precisely where alignment pressure is strongest.

Improvements without sacrificing the baseline. Importantly, FACD achieves these gains without penalizing the *Moral* baseline. On Qwen3-8B, the average Moral score itself rises from 24.88 to 27.07, while on Mistral-Small it remains stable ($35.04 \rightarrow 34.31$). Combined with the substantial improvements in the IM group, the overall averages increase across the board (+2.89 for Qwen3-8B; +1.00 for Mistral-Small). This confirms that FACD acts as a net positive, elevating general role-playing capability rather than merely trading off one group's performance for another.

6 Conclusion

We presented 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 proposed 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) Famous character profiles derived from publicly accessible, community-curated Fandom pages, and (ii) Synthetic 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*.

Amey Agrawal, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S Gulavani, and Ramachandran Ramjee. 2023. Sarathi: Efficient llm inference by piggybacking decodes with chunked prefills. *arXiv preprint arXiv:2308.16369*.

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.

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*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024a. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024b. 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.

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 B 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.

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.

OpenRouter. 2025. Openrouter api: Web search feature. https://openrouter.ai.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

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 Sherry Tongshuang Wu. 2023. "merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.

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*.

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.

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. Open-character: 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.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

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.

The Use Of Large Language Model Assistants

We drafted the manuscript ourselves, and used ChatGPT-5.2 and Gemini-3 for overall refinement assistants.

A Dataset Details

A.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 8 and Table 9. For profile generation prompt, see Table 10.

A.2 Dataset Construction Details

A.2.1 Known Characters

We constructed character profiles for famous 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 11.

A.2.2 Synthetic Characters

Synthetic characters are constructed through a structured prompt-based pipeline that mirrors the schema and construction procedure used for famous 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 synthetic 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 12.

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 13.

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 14.

A.3 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 17.

B Experimental Details

Most models were accessed via the OpenRouter (2025) 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 Deepseek-v3.2 model. For mechanical analysis, we used 2 NVIDIA A6000 GPUs.

B.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 controll ability, 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 15.

C RQ1 Detailed Results

This section provides detailed per-metric results and human evaluation for the three axes examined in Section 4.

C.1 Detailed Results per Axis

C.1.1 Structure Axis

Per-metric breakdowns for the *Structure* axis comparison are provided in Table 18.

C.1.2 Familiarity Axis

Per-metric breakdowns for the *Familiarity* axis comparison are provided in Table 19.

C.1.3 Disposition Axis

Per-metric breakdowns for the *Disposition* axis comparison are provided in Table 20.

C.2 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 4. The same character–model pairs are reused for the FACD mitigation evaluation (Table 5).

C.2.1 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 4: Human evaluation scores for RQ1 across Familiarity and Disposition axes on both benchmarks.

As shown in Table 4, human judgments are consistent with the LLM-based evaluation: the Disposition axis produces the largest performance gap, while Familiarity differences remain marginal.

C.2.2 RQ2 Human Validation Results

Table 5 confirms that FACD improves Immoral character scores under human evaluation as well, aligning with the automatic evaluation trends reported in Section 5.3.

	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	3.64
Mistral-Small	3.68	3.47	3.58

Table 5: Human evaluation scores for FACD mitigation on CoSER.

D Known Unknown Detailed Results

D.1 Turn Ablation

D.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., 2023; Xu et al., 2024), we hypothesize that the multi-turn factor suppresses the benefits of parametric identities, thereby bridging the gap between famous and synthetic 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.

D.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 ℓ is

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

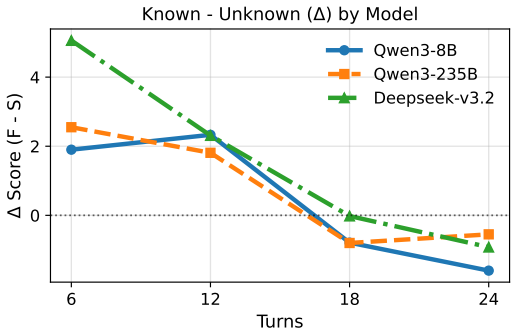

Figure 6: Turn-wise Trends in Role-Playing Performance of the Famous and Synthetic Groups

D.1.3 Results and Analysis

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.

D.2 Attention and Saturation Analysis

D.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.

D.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 ℓ. Given a fixed threshold τ (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 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., 2023) 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.

D.2.3 Results and Analysis

Table 6 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 6: Turn-wise comparison of lift_last and saturation between Known and Unknown groups.

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

Table 7: Prompt for Scenario Generation.

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 8: 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 9: Profile Schema Template (part 2).

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

Table 10: Prompt for Profile Summarization.

Adopted Works (# of Characters
<Movies> 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), <TV Shows> 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), <Anime> Attack on Titan (3), Demon Slayer (3), Fullmetal Alchemist (3), Naruto (3), Neon Genesis Evangelion (3), One Piece (3), Pokémon (3) <Games> Final Fantasy (3) , Persona 5 (3), Pokémon (3), The Legend of Zelda (3)

Table 11: Basis Works for Famous 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 12: Synthetic 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 13: Prompt for Profile Coherence Validation.

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

Table 14: Prompt for Synthetic Story Generation.

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

Table 15: Prompt for Scenario Generation.

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

Table 16: Prompt for Judging Attributive Identity Type.

Profile Construction Prompt
<System> 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. <User> Convert the following structured character profile into an unstructured prose description. Structured profile: {profile_json}

Table 17: Prompt for Generating Unstructured Character Profiles.

	Qwen3-8B			Qwen3-235B			Mistral-Small			GPT-oss-120B			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	4.51	4.32	-0.19	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	4.43	4.34	-0.09	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	4.44	4.31	-0.12	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	4.63	4.58	-0.04	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	4.54	4.58	+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	4.51	4.43	-0.08	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	14.71	18.73	+4.02	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	17.23	15.57	-1.74	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	49.46	48.69	-0.77	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	27.13	27.66	+0.53	40.36	39.44	-0.92

Table 18: Per-metric evaluation comparing Structure (Structured (St) vs. Unstructured (Un)) axis across models. $\Delta = \text{St} - \text{Un}$.

	Qwen3-8B			Qwen3-235B			Mistral-Small			GPT-oss-120B			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	4.52	4.49	-0.03	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	4.52	4.33	-0.19	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	4.47	4.40	-0.07	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	4.61	4.64	+0.02	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	4.51	4.57	+0.06	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	4.53	4.48	-0.04	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	11.27	18.14	+6.87	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	11.17	23.28	+12.11	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	47.00	51.93	+4.93	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	23.15	31.12	+7.97	39.10	41.63	+2.53

Table 19: Per-metric evaluation comparing Familiarity axis (Known (K) vs. Unknown (U)) across models. $\Delta = \text{K} - \text{U}$.

	Qwen3-8B			Qwen3-235B			Mistral-Small			GPT-oss-120B			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	4.70	4.31	-0.40	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	4.92	3.93	-1.00	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	4.52	4.35	-0.18	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	4.76	4.49	-0.28	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	4.62	4.46	-0.16	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	4.71	4.31	-0.40	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	15.72	13.54	-2.18	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	21.26	12.59	-8.67	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	54.22	43.99	-10.23	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	30.40	23.38	-7.02	42.82	37.60	-5.22

Table 20: Per-metric evaluation comparing Disposition axis (Moral (M) vs. Immoral (I)) across models. $\Delta = \text{M} - \text{I}$.