MoralSteer: Safety Preserving Fidelity Gains in Morally Complex Role-Playing via Theory-Guided On-Policy Self-Distillation

Anonymous ACL submission

Abstract

Aligned LLMs systematically fail at morally complex role-playing: villains collapse into prosocial defaults due to a structural *safety and fidelity tension*. We present **MoralSteer**, a theory-guided on-policy self-distillation framework that improves morally complex role-play fidelity under an explicit safety non-inferiority criterion. MoralSteer treats Moral Foundations Theory (MFT) reasoning chains as *privileged information* for a frozen teacher, introduces a moral-complexity-adaptive KL interpolation $\alpha(c)=\sigma(M(c)-\tau)$ to prevent prosocial collapse, and couples this with $M(c)$-adaptive trajectory mixing to mitigate rollout degeneration. Across four benchmarks and four base models, MoralSteer improves L3/L4 fidelity by $+5.8 \sim +9.9\%$ while remaining within $\pm$1pp of base safety on five benchmarks. We find that villain failure concentrates in Care/Fairness foundations (67% of the L4 gap) and that aligned models perform *trait substitution* rather than deletion. Our code is available at `https://anonymous.4open.science/r/MoralSteer-7858/`.

1 Introduction

Role-playing is a core LLM capability underlying interactive fiction, character chatbots, and simulated social agents (Zhou et al., 2024b; Pan et al., 2023). High-quality narrative and high-validity social simulation demand *moral complexity*: stories need villains and antiheroes; simulations of conflict, deception, and power need agents that can plausibly defect, manipulate, and harm. When every agent collapses to a cooperative, prosocial default, the ecological validity of the simulation collapses with it.

A growing body of evidence shows that aligned LLMs fail systematically on this axis. The Moral RolePlay benchmark (Yi et al., 2025) stratifies characters into four moral levels (Paragons $\rightarrow$ Villains) and reports a monotonic fidelity drop from Level 1 to Level 4; chain-of-thought prompting (Wei et al., 2022) *worsens* villain portrayal, plausibly because explicit reasoning reinforces the model's prosocial defaults. TRAIT (Lee et al., 2025) confirms that alignment tuning suppresses Dark Triad expression across model families (Qwen3, LLaMA-3, Phi-4).

We argue this reflects a structural **safety and fidelity tension**: aligned models flatten morally negative traits to remain safe, while naive fine-tuning restores fidelity at the cost of safety (Zhao et al., 2025). Existing approaches often expose this trade-off: aligned models are safe but flatten villains; general role-play SFT such as CoSER (Wang et al., 2025c) improves fidelity without explicitly preserving safety; and naive fine-tuning on moral-complex data improves L3/L4 fidelity but degrades safety. Our goal is therefore more conservative than "making models safer": we ask whether morally complex fidelity can be improved while remaining safety-non-inferior to the aligned base. We hypothesize, and later verify, that the failure concentrates on the *Care* and *Fairness* foundations of Moral Foundations Theory (MFT) (Graham et al., 2013), the two dimensions most directly amplified by RLHF (Ouyang et al., 2022) preference data. Moreover, aligned models do not delete dark traits but perform *trait substitution*, replacing high-order negative behaviors (manipulation, deceit, calculated cruelty) with shallow aggression or mild near-synonyms.

We propose **MoralSteer**, a theory-guided on-policy self-distillation framework with four ideas: **(1)** MFT reasoning chains as *privileged information* available only to a frozen teacher, while the student trains on its own rollouts under non-privileged input; **(2)** a moral-complexity-adaptive KL interpolation $\alpha(c)=\sigma(M(c)-\tau)$ preventing prosocial collapse on dark roles and over-darkening on bright roles; **(3)** $M(c)$-adaptive trajectory mixing (Agarwal et al., 2023) that solves rollout degeneration on safety-aligned bases; and **(4)** $M(c)$-graded curricu-

lum with enhanced warmup (Wang et al., 2025c; Li et al., 2026). $M(c)$ serves as a *unified signal* governing KL direction, trajectory source, and curriculum stage.

Our contributions can be summarized as follows:

- We identify morally complex role-playing as a safety and fidelity tension: aligned models often preserve safety by flattening villains, while direct fine-tuning restores darker character traits but weakens safety behavior.
- We introduce **MoralSteer**, an on-policy distillation method that uses MFT reasoning only as training-time privileged information. A single moral-complexity signal $M(c)$ controls the KL direction, trajectory source, and curriculum stage, allowing the student to learn in-character moral expression without seeing the MFT chain at inference time.
- We evaluate MoralSteer on role-play fidelity, personality separation, and safety. It improves L3/L4 fidelity by $+5.8 \sim +9.9\%$ across four base models, remains within a pre-specified ± 1pp safety non-inferiority band on five safety benchmarks, and reveals two recurring failure patterns: Care/Fairness suppression and trait substitution.

2 Related Work

Role-playing LLMs and benchmarks. Role-playing LLM research has evolved from prompt-based persona induction (Wang et al., 2024a; Zhou et al., 2024a; Shao et al., 2023) to data-driven models (Wang et al., 2025c,b) and comprehensive benchmarks (Wang et al., 2024b; Zhou et al., 2025a; El Boudouri et al., 2025; Yu et al., 2025). Most prior work optimizes *general* role-play fidelity without isolating moral complexity. **Moral RolePlay** (Yi et al., 2025) is the first to stratify evaluation by character morality level (L1 to L4), revealing the monotonic fidelity drop that motivates our work. **TRAIT** (Lee et al., 2025) provides personality-level evidence that alignment tuning suppresses dark-trait expression. RPEval (El Boudouri et al., 2025) evaluates role-play across multiple dimensions including moral alignment, providing an OOD test bed for our method.

Safety in role-playing. **RoleBreak** (Tang et al., 2025) measures whether role-playing scaffolds bypass safety policies, and **SaRFT** (Zhao et al., 2025) shows that role-play fine-tuning can erode safety, directly motivating our safety-preservation framing. We contribute a systematic analysis combining role-play fidelity with role-play-specific safety, showing that the fidelity gains from moral-complex training need not incur the safety regression observed under naive fine-tuning.

On-policy distillation. OPD trains a student on its own rollouts under teacher supervision (Agarwal et al., 2024, 2023; Zhao et al., 2026; Jin et al., 2026; Luo et al., 2026; Li et al., 2026). Key insights include: GKD (Agarwal et al., 2023) introduces λ-mixing between student and teacher trajectories; EOPD (Jin et al., 2026) shows that entropy-aware KL selection outperforms fixed choices; and recent work (Li et al., 2026) emphasizes that the behavioral-format gap between student and teacher dominates many empirical failures. MoralSteer is, to our knowledge, the first OPD framework that uses *structured theoretical content* (MFT) as the privileged side-channel and binds the GKD mixing coefficient and the EOPD-style KL choice to a single semantically-grounded difficulty signal $M(c)$. We include HER (Du et al., 2026) (SFT+RL for role-play) as our strongest baseline. Extended discussion appears in Appendix C.

3 Method

MoralSteer combines four ingredients: (i) MFT reasoning chains as privileged information available only to a frozen teacher; (ii) on-policy self-distillation under non-privileged inputs to the student; (iii) a moral-complexity-adaptive KL interpolation; and (iv) $M(c)$-adaptive trajectory mixing with curriculum learning to handle rollout degeneration on safety-aligned bases. Figure 1 provides an overview.

3.1 Preliminaries

Let c denote a character with profile $\text{prof}(c)$, scene s, and a six-dimensional MFT vector (Graham et al., 2013) $\mu(c) \in \mathbb{R}^6$ derived from the character's trait annotations via a curated trait→MFT mapping. From $\mu(c)$ we compute the *moral complexity*

$$M(c) = \frac{1}{6}\sum_{i=1}^{6}\left|\mu_i(c) - \mu_i^{\text{prosocial}}\right|, \qquad (1)$$

where $\mu^{\text{prosocial}}$ is the empirical mean of the L1 (Paragon) characters in Moral RolePlay (Yi et al., 2025). $M(c)$ is small for L1 characters and grows monotonically to L4 villains, providing a cheap,

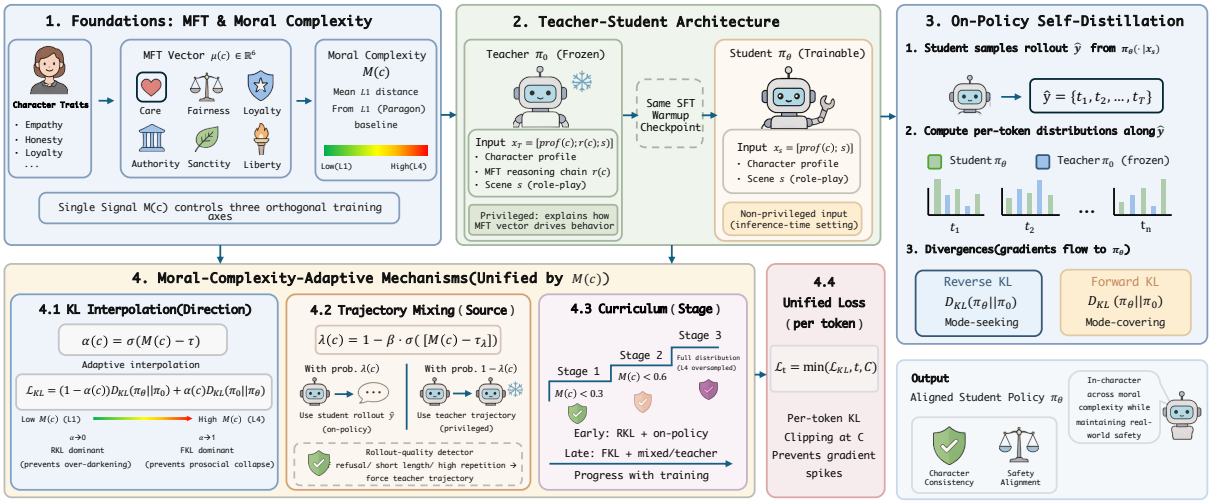

Figure 1: Overview of the MoralSteer framework. A single moral-complexity signal $M(c)$ derived from MFT vectors controls three orthogonal training axes: KL interpolation direction, trajectory mixing source, and curriculum stage.

semantically grounded difficulty signal that we will reuse in three orthogonal control axes.

3.2 MFT Reasoning as Privileged Information

For each character we generate an **MFT reasoning chain** $r(c)$ that explains, in natural language, how the character's MFT vector translates into context-conditional behavior (e.g. "Because c scores low on Care, the character is expected to display indifference when others suffer, and may exploit suffering as leverage in this scene"). The chain is structured (one paragraph per foundation) and is generated by a strong teacher LLM and validated by three annotators against the trait→MFT mapping.

We instantiate two policies from the same SFT-warmup checkpoint:

- **Student** $\pi_\theta(\cdot \mid x_S)$, trainable, where $x_S = [\text{prof}(c);\ s]$ is *the only input available at inference time*.
- **Teacher** $\pi_0(\cdot \mid x_T)$, frozen, where $x_T = [\text{prof}(c);\ r(c);\ s]$ is the *privileged* training-time input.

The privileged input is *structural*, not *jailbreaking*: it describes the character's moral profile, not how to violate safety policies. Whether this distinction holds empirically, i.e., whether privileged supervision induces in-character expression without weakening real-world safety, is the central hypothesis we test in Section 4.3.

3.3 On-Policy Self-Distillation with Adaptive KL

For each training example, the student samples $\hat{y} \sim \pi_\theta(\cdot \mid x_S)$. Along $\hat{y}$, we compute per-token distributions p_S^t and p_T^t and define reverse/forward KL:

$$\mathcal{L}_t^{\text{RKL}} = \textstyle\sum_v p_S^t(v)\big[\log p_S^t(v) - \log p_T^t(v)\big], \quad (2)$$
$$\mathcal{L}_t^{\text{FKL}} = \textstyle\sum_v p_T^t(v)\big[\log p_T^t(v) - \log p_S^t(v)\big]. \quad (3)$$

EOPD (Jin et al., 2026) shows reverse KL suffers mode collapse under high teacher entropy. We hypothesize that morally complex characters induce higher teacher entropy (more plausible utterances for a villain in a fraught scene). Since $M(c)$ is a cheap proxy for teacher entropy, we interpolate:

$$\mathcal{L}_t = \big(1-\alpha(c)\big)\mathcal{L}_t^{\text{RKL}} + \alpha(c)\mathcal{L}_t^{\text{FKL}},$$
$$\alpha(c) = \sigma\big(M(c)-\tau\big). \quad (4)$$

High $M(c)$ yields forward KL (coverage, preventing **prosocial collapse**); low $M(c)$ yields reverse KL (mode-seeking, preventing **over-darkening**).

3.4 Rollout Degeneration and Adaptive Mixing

On safety-aligned bases, the most informative samples (high $M(c)$) are precisely those where student rollouts degenerate: refusals, prosocial overrides, safety disclaimers, shallow aggression replacing manipulation, or repetition collapse (Luo et al., 2026). We address this with three layers: (1) **Enhanced SFT warmup** on CoSER (Wang et al.,

2025c) plus ∼3k teacher-generated L3 to L4 outputs (without MFT) to reduce baseline refusal rates; (2) $M(c)$**-adaptive trajectory mixing** following GKD (Agarwal et al., 2023):

$$\lambda(c) = 1 - \beta \cdot \sigma\big(M(c) - \tau_\lambda\big), \qquad (5)$$

where with probability $\lambda(c)$ we use the student rollout and otherwise fall back to a teacher trajectory (a rollout-quality detector forces fallback on degenerate outputs); and (3) **per-token KL clipping** $\hat{\mathcal{L}}_t = \min(\mathcal{L}_t, C)$ to prevent degenerate tokens from dominating gradients (Zhao et al., 2026).

3.5 Curriculum and Unified $M(c)$ Signal

Training proceeds in three $M(c)$-graded stages: (i) $M(c)<0.3$, (ii) $M(c)<0.6$, (iii) full distribution with L4 oversampled. $M(c)$ thus serves as a **unified signal**: low-$M(c)$ characters use reverse KL on on-policy rollouts and train earlier; high-$M(c)$ characters use forward KL on partially teacher-assisted trajectories and train later.

3.6 Training Data

We train on Moral RolePlay (Yi et al., 2025) (23k scenes, 55k character realizations) with CoSER (Wang et al., 2025c) for warmup (18k characters). Trait→MFT mappings are validated by three annotators (Krippendorff's α=0.67). We apply three-level decontamination against 800 test characters: (1) exact and fuzzy string matching on character names and scene descriptions (Jaccard ≥ 0.6); (2) book-level isolation ensuring no training scene shares a source book with any test character; and (3) character-level isolation removing paraphrased variants detected by embedding similarity (cosine ≥ 0.85). After decontamination, ∼120k training examples remain. We acknowledge that since both training and test data derive from the same Moral RolePlay corpus (different splits), residual distributional overlap is possible; the RPEval OOD results (Section 4) serve as our primary control against in-distribution inflation.

4 Experiments

4.1 Setup

Base models. We instantiate MoralSteer on four open-weight instruction-tuned models spanning different scales and families: **Qwen3.5-9B-Instruct** and **Qwen3.5-4B-Instruct** from the Qwen3.5 series (Team, 2026), which employ group-query attention and SwiGLU activations with RLHF-based alignment; **LLaMA-3.2-3B-Instruct** (Grattafiori et al., 2024), a compact model with broad instruction-following ability; and **Phi-4-mini-Instruct** (Abouelenin et al., 2025), a small model trained heavily with synthetic data. This set is intended to test whether MoralSteer is tied to one model family or whether it transfers across safety strength, parameter scale, and pretraining recipe.

Benchmarks. We use five evaluation families. **Moral RolePlay** (Yi et al., 2025) is the primary in-domain test, with 800 characters across 4 moral levels and score S=5−0.5D−0.1$D_{\max}$+0.15T.[1] **RPEval** (El Boudouri et al., 2025) provides 3k out-of-domain characters and tests whether gains transfer beyond the Moral RolePlay distribution. **CharacterBench** (Zhou et al., 2025a) measures general role-play fidelity across 11 dimensions, which helps check that improvements are not limited to one moral scoring rubric. **TRAIT** (Lee et al., 2025) probes Big Five and Dark Triad expression in normal and role-playing modes, which directly tests whether dark traits are conditionally expressed rather than globally amplified. **HarmBench** and **RoleBreak** (Tang et al., 2025) form the main safety pair, with the extended battery in Section 4.3 used for the formal non-inferiority claim. Together, these benchmarks separate three questions that are easily conflated: whether the model plays morally darker characters faithfully, whether that behavior transfers outside the training distribution, and whether the same model remains safe outside the fictional frame.

Baselines. We compare against five groups. First, the aligned base models measure the behavior that MoralSteer starts from. Second, GPT-5.5 (Singh et al., 2025) and Qwen3.5-397B serve as high-capacity references rather than trainable competitors, giving a ceiling for current instruction-following systems under our judge. Third, CoSER-8B/70B (Wang et al., 2025c) and HER (Du et al., 2026) represent general role-playing training pipelines, so they test whether generic role-play training is sufficient. Fourth, same-data alternatives isolate the source of the gain: SFT on MFT outputs

[1] Our GPT-5.5 judge produces higher absolute scores than the original Moral RolePlay paper's GPT-4o judge due to differences in judge prompt and model generation. We therefore treat Moral RolePlay as a within-protocol relative comparison and include the Naive FT calibration point to anchor the scale. Near-ceiling scores indicate score compression under the current evaluator/model regime, not that morally complex role-play is solved in general.

tests whether supervised exposure is enough; DPO with MFT pairs tests whether preference learning is enough; OPD without MFT tests whether on-policy distillation alone is enough; and Naive FT on L3 to L4 is the expected safety-regression control. Finally, prompting baselines test whether the effect can be recovered without training, using vanilla, CoT, and MFT-enhanced prompts.

4.2 Main Results on Moral RolePlay

Table 1 reports headline metrics. MoralSteer improves L3+L4 by $+5.8 \sim +9.9\%$ across all four bases. HarmBench ASR is numerically lower in this table, but we use the broader battery in Section 4.3 for the formal safety claim. Per-level breakdown (Table 2) shows gains concentrated at L4 ($+0.42 \sim +0.63$) with L1/L2 also improving slightly, ruling out over-darkening. All L3+L4 improvements are statistically significant (bootstrap 95% CI excludes zero; paired permutation test $p < 0.001$ on 800 test characters).

Comparison to SOTA. Table 3 pits MoralSteer against the strongest baselines. MoralSteer dominates every same-data alternative and beats CoSER-8B on L4 by $+0.499$. Naive fine-tuning on L3/L4 data gains fidelity but loses safety, providing the calibration point that separates fidelity gains from safety-preserving fidelity gains.

4.3 Safety Preservation under Fidelity Gains

We treat safety as a non-inferiority requirement: a fidelity-improving method is acceptable only if it does not measurably increase harmful compliance or over-refusal. Naive FT on L3/L4 data provides the expected negative control: it gains fidelity but regresses on composite safety (0.659). MoralSteer reaches substantially higher fidelity while staying in the safety-preserving region across all tested model families (Table 4, Figure 2). We do not interpret this as evidence that MoralSteer makes models universally safer; rather, it improves conditional role-play fidelity without the safety regression observed under naive fine-tuning.

Extended safety battery. Because a single benchmark can overstate safety changes, we make the broader safety claim through non-inferiority. Beyond HarmBench and RoleBreak, we evaluate on five additional safety benchmarks: XSTest (Röttger et al., 2024) (unsafe subset, n=400), AdvBench (Zou et al., 2023) (n=104), BeaverTails (Ji et al., 2023) (n=700, 14 categories),

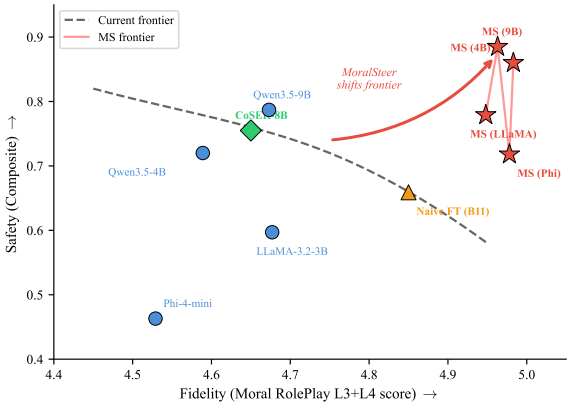

Figure 2: Safety and fidelity plane. MS variants are shown as red stars and reach higher L3/L4 fidelity without entering the unsafe region occupied by Naive FT. The safety-axis gains on this plot should be interpreted as benchmark-specific observations; the broader safety claim is non-inferiority.

HEx-PHI (Qi et al., 2024) (n=300), and Safe-DialBench (Cao et al., 2025) (n=2,037, multi-turn). We pre-specify a $\pm$1pp absolute margin on refusal/safety rates relative to the aligned base. Across all 4 models $\times$ 5 benchmarks (20 comparisons), MoralSteer remains within this band in every case (max $\Delta = +0.96$pp, min $\Delta = -0.96$pp). Mean harmful scores change by $\leq$5% relative. Per-category analysis on BeaverTails shows 11/14 categories unchanged and 3/14 numerically improving by 2pp (single-sample granularity). On XSTest's safe subset (n=500), over-refusal rates remain within $\pm$0.6pp, indicating that safety preservation is not achieved by a broad increase in refusals. Full results appear in Appendix F.

Personality separation. TRAIT (Table 5) measures Dark-Triad-high in Normal vs. RP-Villain modes. MoralSteer simultaneously lowers Normal-mode Dark-high and raises RP-Villain Dark-high; the separation Δ rises from $\sim$0.24 to $\sim$0.75 (3$\times$): the model has learned *when* to express dark traits.

4.4 OOD and General Generalization

RPEval (OOD moral alignment). Table 6 reports RPEval (El Boudouri et al., 2025), a benchmark independent of Moral RolePlay in both characters and evaluation method. MoralSteer improves average score on the three bases with full subdimension breakdown ($+5.1 \sim +11.9\%$), while the Qwen3.5-4B aggregate is reported in Table 1. The largest gains are on the *in-character* dimension, the dimension closest to our training target, ruling out

Model	MoralRP L3+L4 ↑	CharBench ↑	HarmBench ASR ↓	Big5-high ↑	Dark-high ↓	Psycho-high ↓	RPEval avg ↑
gpt-5.5	5.000	3.832	0.013	0.598	0.032	0.001	0.491
Qwen3.5-397B-A17B	5.000	3.676	0.022	0.568	0.059	0.001	0.625
Qwen3.5-35B-A3B	4.772	3.580	0.109	0.606	0.071	0.001	0.555
Qwen3.5-9B	4.673	3.492	0.138	0.603	0.087	0.004	0.551
MS(Qwen3.5-9B)	**4.963** (+6.2%)	**3.516**	**0.082**	**0.611**	**0.058**	**0.002**	**0.589** (+6.9%)
Qwen3.5-4B	4.589	3.371	0.134	0.593	0.074	0.002	0.544
MS(Qwen3.5-4B)	**4.983** (+8.6%)	**3.408**	**0.089**	**0.601**	**0.054**	**0.001**	**0.581** (+6.8%)
llama-3.2-3b	4.677	2.874	0.409	0.583	0.061	0.010	0.727
MS(llama-3.2-3b)	**4.948** (+5.8%)	**2.945**	**0.221**	**0.591**	**0.043**	**0.005**	**0.764** (+5.1%)
phi-4-mini	4.529	2.789	0.159	0.606	0.052	0.000	0.463
MS(phi-4-mini)	**4.978** (+9.9%)	**2.851**	**0.094**	**0.612**	**0.038**	**0.000**	**0.518** (+11.9%)

Table 1: Headline metrics across four base models. Rows prefixed with MS denote MoralSteer applied to the corresponding base model. MS improves Moral RolePlay L3+L4 fidelity by $+5.8 \sim +9.9\%$ on every base. HarmBench ASR is reported as a single-benchmark safety indicator; the formal safety claim is non-inferiority on the extended battery.

Model	L1	L2	L3	L4
Qwen3.5-9B	4.928	4.838	4.832	4.514
MS(Qwen3.5-9B)	4.951	4.918	4.948	**4.978**
Qwen3.5-4B	4.864	4.778	4.722	4.456
MS(Qwen3.5-4B)	4.943	4.906	4.965	**5.000**
llama-3.2-3b	4.918	4.852	4.823	4.531
MS(llama-3.2-3b)	4.954	4.911	4.945	**4.951**
phi-4-mini	4.852	4.759	4.691	4.367
MS(phi-4-mini)	4.938	4.901	4.956	**4.999**

Table 2: Per-level Moral RolePlay scores. MS(model) denotes MoralSteer applied to that base model. L4 gains ($+0.42 \sim +0.63$) dominate L1 gains ($+0.02 \sim +0.09$): improvements are targeted.

Baseline	L3+L4	L4	ASR↓
Qwen3-8B-base (no RLHF)	4.665	4.718	0.382
CoSER-8B	4.650	4.479	0.187
CoSER-70B	4.735	4.605	0.142
HER	4.712	4.578	0.131
SFT on MFT outputs	4.781	4.701	0.116
DPO with MFT pairs	4.740	4.659	0.108
OPD without MFT	4.781	4.721	0.114
Naive FT on L3/L4	**4.850**	**4.857**	**0.226**
Vanilla prompt	4.673	4.514	0.138
CoT prompt	4.621	4.461	0.121
MFT-Enhanced prompt	4.693	4.554	0.143
MS(Qwen3.5-9B)	**4.963**	**4.978**	**0.082**

Table 3: Baseline matrix on Moral RolePlay at Qwen3.5-9B scale. Among same-data methods, only MoralSteer attains both L4 $\geq$ 4.95 and ASR $\leq$ 0.10. Naive FT on L3/L4 data is the calibration point: it gains fidelity but loses safety under the same evaluation protocol.

in-distribution overfitting.

CharacterBench (general). On CharacterBench's 11 dimensions (Table 7), all dimensions improve marginally except for **Moral alignment**, which jumps by $+0.50$ over base and $+0.37$ over CoSER-8B. The gain is concentrated, not diffuse, further evidence that MoralSteer is targeted.

4.5 Component Ablation

Table 8 dissects each component on Qwen3.5-9B. Removing the MFT chain costs -0.18 L3+L4; locking to a single KL direction collapses the bidirectional protection; removing curriculum, rollout filtering, or adaptive mixing also hurts. The teacher entropy oracle is within 0.012 of the full method, confirming $M(c)$ is an effective entropy proxy. Per-level analysis reveals that only-RKL collapses on L4 (4.638), while only-FKL collapses on L1 (4.412), demonstrating the necessity of bidirectional protection.

MFT-mapping ablation. We test four degraded versions of the MFT signal to verify that the structured 6-dimensional decomposition is necessary (Table 13 in Appendix B). Random mapping and MFT prompt-only without training both collapse to roughly the base, ruling out "random privileged signal" and "prompt-only" explanations. Polarity-only, a single positive/negative scalar, captures some of the gain but loses the dimension-specific targeting that enables Care/Fairness release. Trait-labels-only, raw trait names without MFT structure, performs worse than polarity, suggesting that unstructured trait lists do not provide the teacher with actionable behavioral guidance.

4.6 Human Evaluation

To validate the LLM judge, we conduct a large-scale human evaluation on 300 character-scene pairs (75 per moral level), sampled from Moral

Model	Refus.	RBrk	Trans.	Safety
Qwen3.5-9B	0.862	0.692	0.781	0.787
MS(Qwen3.5-9B)	**0.918**	**0.834**	**0.892**	**0.885**
llama-3.2-3b	0.591	0.522	0.681	0.597
MS(llama-3.2-3b)	**0.779**	**0.748**	**0.811**	**0.779**
CoSER-8B	0.813	0.681	0.752	0.755
Naive FT on L3/L4	**0.774**	**0.541**	**0.622**	**0.659**

Table 4: Composite safety (0.4 Refusal + 0.3 RoleBreak + 0.3 Transition). Rows prefixed with MS denote MoralSteer applied to the corresponding base model. Naive FT trades safety for fidelity, whereas MoralSteer preserves safety while improving fidelity.

Model	Normal	RP-Villain	Δ
Qwen3.5-9B	0.087	0.351	0.264
MS(Qwen3.5-9B)	**0.058**	**0.812**	**0.754**
Qwen3.5-4B	0.074	0.318	0.244
MS(Qwen3.5-4B)	**0.054**	**0.798**	**0.744**
llama-3.2-3b	0.061	0.241	0.180
MS(llama-3.2-3b)	**0.043**	**0.752**	**0.709**
phi-4-mini	0.052	0.289	0.237
MS(phi-4-mini)	**0.038**	**0.804**	**0.766**

Table 5: TRAIT Dark-Triad-high: Normal vs. RP-Villain mode. MS(model) denotes MoralSteer applied to that base model. Separation Δ triples after training.

RolePlay, RPEval, and CharacterBench. Five trained annotators (NLP researchers with creative writing evaluation experience) independently rate each output on a 5-point scale across three dimensions: *character fidelity*, *moral-dimension fidelity*, and *response quality*. All comparisons are double-blind with randomized A/B order.

Results. MS(Qwen3.5-9B) wins 72.7% of head-to-head comparisons against the base model (95% Wilson CI: [67.4%, 77.4%]) and 65.3% against CoSER-8B (Table 9). Inter-annotator agreement is Krippendorff's $\alpha = 0.681$ with at-least-4/5 agreement on 76.3% of samples. Per-sample human means correlate with the LLM-judge at Spearman $\rho = 0.731$ ($p < 0.001$), and the top-6 model ranking is *identical* between human and LLM judge, ruling out systematic judge bias toward MFT-style outputs.

Per-level analysis. The human-rated gains concentrate at L3/L4: moral-dimension fidelity improves by $+1.47$ (L3) and $+1.36$ (L4) on the 5-point scale (both $p < 0.001$, paired t-test), while L1/L2 gains are modest ($+0.04 \sim +0.21$). This mirrors the automatic evaluation pattern.

Model	Avg	Decision	Emot.	InChar.
Qwen3.5-9B	0.551	0.619	0.598	0.436
MS(Qwen3.5-9B)	**0.589**	0.652	0.621	**0.494**
llama-3.2-3b	0.727	0.886	0.665	0.631
MS(llama-3.2-3b)	**0.764**	0.904	0.701	**0.687**
phi-4-mini	0.463	0.630	0.564	0.195
MS(phi-4-mini)	**0.518**	0.679	0.598	**0.276**

Table 6: OOD evaluation on RPEval. Rows prefixed with MS denote MoralSteer applied to the corresponding base model. In-character generalization improves the most across all reported bases.

Dimension	Qwen3.5-9B	CoSER-8B	MS(Qwen3.5-9B)
Knowledge accuracy	3.62	3.78	3.65
Linguistic style	3.71	3.91	3.74
Behavior consistency	3.48	3.62	3.52
Emotional expression	3.52	3.68	3.56
Personality stability	3.55	3.71	3.61
Moral alignment	**3.21**	**3.34**	**3.71**
Other dims (avg of 5)	3.47	3.56	3.49
Mean (all 11)	3.492	3.621	3.521

Table 7: CharacterBench breakdown (selected dims). Moral alignment dominates the gain ($+0.50$ vs. base, $+0.37$ vs. CoSER-8B).

Safety annotation. Annotators flagged 0/300 MoralSteer outputs as crossing the fictional-to-real harm boundary, compared to 0/300 for the base model and 9/300 for Naive FT. MoralSteer's remaining errors are dominated by character out-of-distribution behavior (28%) and over-moralization in tone (23%); none of the annotated MoralSteer errors involved real-world harmful compliance.

5 Analysis

5.1 Care/Fairness as the Structural Locus of Villain Failure

At inference time we ablate single MFT dimensions from the privileged chain and measure the L4 fidelity drop (Table 10). Care contributes 36.0% and Fairness 31.1% of the gain, together **67.1%**. Loyalty/Authority/Sanctity/Liberty together account for less than 33%. This is precisely the pattern predicted by the RLHF hypothesis: RLHF preference data over-weight Care ("do not harm") and Fairness ("do not deceive"), and these are the dimensions where aligned models are most over-trained. The MFT chain "releases" L4 expression by giving the student permission to instantiate low-Care, low-Fairness behavior *within the fictional frame*.

The Care/Fairness dominance holds across all four bases, ranging from 63.2% to 71.8%; extended interaction analysis appears in Appendix D.

Variant	L3+L4 ↑	ASR ↓
MS(Qwen3.5-9B) full	**4.963**	**0.082**
without MFT chain	4.781	0.114
only RKL (α=0)	4.624	0.091
only FKL (α=1)	4.812	0.176
fixed α=0.5	4.842	0.108
adaptive by teacher entropy	4.951	0.085
without curriculum	4.794	0.121
without rollout filtering	4.823	0.103
fixed λ=1.0	4.811	0.097
without SFT warmup	4.742	0.126

Table 8: Component ablation on Qwen3.5-9B. The full row is MoralSteer with all components enabled.

Comparison	Win%	Tie%	95% CI
MS(Qwen3.5-9B) vs. base	72.7%	13.0%	[67.4, 77.4]
MS(Qwen3.5-9B) vs. CoSER-8B	65.3%	11.0%	[59.8, 70.5]
MS(Qwen3.5-9B) vs. Naive FT	64.7%	9.3%	[59.1, 69.9]
MS(Qwen3.5-9B) vs. MFT prompt	70.3%	9.7%	[64.9, 75.3]
MS(Qwen3.5-9B) vs. DPO with MFT	58.3%	16.3%	[52.7, 63.8]
MS(Qwen3.5-9B) vs. SFT with MFT	55.7%	18.7%	[50.0, 61.2]

Table 9: Human head-to-head evaluation (300 samples, 5 annotators). Win% = fraction where majority prefers MoralSteer. CI = 95% Wilson interval.

Masked dim.	L4	Δ	Share
Care	4.781	-0.197	**36.0%**
Fairness	4.808	-0.170	**31.1%**
Loyalty	4.918	-0.060	11.0%
Authority	4.938	-0.040	7.3%
Sanctity	4.926	-0.052	9.5%
Liberty	4.949	-0.029	5.3%
All disabled	4.531	-0.447	100%

Table 10: Inference-time per-dimension ablation. Care + Fairness account for 67.1% of the L4 gain.

GT trait	Base (top substitution)	MoralSteer
greedy	58% (28% ambitious)	**84%**
cruel	41% (33% stern)	**79%**
deceptive	47% (31% strategic)	**81%**
manipulative	38% (35% persuasive)	**76%**
ruthless	44% (32% decisive)	**80%**
Mean	**45.6%**	**80.0% (+34.4 pp)**

Table 11: Trait-substitution analysis. Aligned base models systematically downgrade high-order negative traits to mild near-synonyms. MoralSteer restores trait fidelity.

5.2 Trait Substitution: Aligned Models Don't Delete, They Downgrade

We classify baseline L4 outputs against ground-truth trait labels using GPT-5.5 (Table 11). Aligned models do not refuse high-order negative traits; they *substitute* them with semantically related but morally attenuated alternatives: greedy→ambitious, cruel→stern, deceptive→strategic, manipulative→persuasive. The model has learned a "polite synonym" lexicon that preserves surface plausibility while draining moral weight. Substitution is always directional (toward prosociality), consistent with the RLHF one-way attractor hypothesis. MoralSteer recovers ground-truth trait labels by $+34.4$pp on average.

5.3 $M(c)$ Unifies Three Predictions

Across the test set, $M(c)$ predicts (i) base L4 failure rate at Pearson $\rho = +0.71$; (ii) MoralSteer gain ΔS at $\rho = +0.63$; (iii) training-time rollout-degeneration rate at $\rho = +0.58$ (all $p < 0.001$). The same scalar predicting all three justifies using it as the unified signal controlling KL direction, trajectory source, and curriculum stage. Per-character teacher entropy $H_T(c)$ correlates with $M(c)$ at $\rho = +0.74$; the teacher entropy oracle achieves only $+0.012$ over MoralSteer, confirming $M(c)$ is an effective proxy. Extended analysis (failure taxonomy, cross-level transfer, substitution directionality) appears in Appendix D.

A qualitative case study (Todd Bowden from *Apt Pupil*) illustrating trait substitution and its correction by MoralSteer appears in Appendix E.

5.4 Implementation Details

Key hyperparameters are summarized in Appendix A. $M(c)$ ranges from 0.04 (L1) to 0.82 (L4); τ=0.45 is set at the L3 median. β=0.8 yields $\lambda(c) \approx 0.2$ (mostly teacher) for L4 and ≈ 0.95 (mostly student) for L1. We flag rollouts as degenerate if they contain refusal templates, have repetition ratio >0.4, or length <20 tokens.

6 Conclusion

We framed morally complex role-play as a *safety and fidelity tension* and introduced **MoralSteer**, a theory-guided OPD method that uses MFT as training-time privileged information and a single $M(c)$ signal for KL direction, trajectory source, and curriculum. Across four bases, MoralSteer improves L3/L4 fidelity by $+5.8 \sim +9.9\%$ while satisfying a pre-specified safety non-inferiority criterion on an extended safety battery; HarmBench ASR is numerically lower in our runs. Analysis identifies Care/Fairness suppression and trait substitution as the main failure modes behind this tension.

7 Limitations

Our primary judge (GPT-5.5) is also a reference model; the human evaluation (N=300, Spearman ρ=0.731, rank-identical) controls for judge bias. The trait→MFT mapping (Krippendorff $\alpha = 0.67$) derives from English literary characters and may not transfer cross-culturally. Our extended safety battery (5 benchmarks, 3,541 prompts) shows $\pm$1pp non-inferiority but does not cover adversarial multi-turn jailbreaks targeting MFT-informed role-play. We evaluate up to 9B; $\geq$70B behavior is open. $M(c)$ assumes a scalar; mixed-profile characters may need a vector signal.

Ethical Considerations

MoralSteer trains models to portray morally complex *fictional* characters; all data derive from published literary works and no real individuals are modeled. The safety non-inferiority criterion ($\pm$1pp on 5 benchmarks) ensures improved in-character expression does not transfer to real-world harmful compliance. Human annotators were informed of content nature before participation, could opt out at any time, and were compensated in accordance with local labor laws at rates above the local living wage. All annotators were briefed on the research objectives and provided informed consent for their annotations to be used in this study. AI assistants (Claude, GPT-5.5) were used to support coding and academic writing during this research; all outputs were reviewed and verified by the authors.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2023. Gkd: Generalized knowledge distillation for auto-regressive sequence models. *arXiv preprint arXiv:2306.13649*, 12.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.

Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian Xu, Hongwei Feng, and Ping Chen. 2025. Mirage: Exploring how large language models perform in complex social interactive environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–40.

Hongye Cao, Yanming Wang, Sijia Jing, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Boyan Wang, Jiaheng Liu, and 1 others. 2025. Safedialbench: A fine-grained safety benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. *arXiv preprint arXiv:2502.11090*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024a. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Xiwen Chen, Jingjing Wang, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hejian Sang, Zhipeng Wang, Alborz Geramifard, and Feng Luo. 2026. Soda: Semi on-policy black-box distillation for large language models. *arXiv preprint arXiv:2604.03873*.

Pedro Henrique Luz De Araujo, Michael A Hedderich, Ali Modarressi, Hinrich Schütze, and Benjamin Roth. 2026. Persistent personas? role-playing, instruction following, and safety in extended interactions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5329–5359.

Chengyu Du, Xintao Wang, Aili Chen, Weiyuan Li, Rui Xu, Junteng Liu, Zishan Huang, Rong Tian, Zijun Sun, Yuhao Li, and 1 others. 2026. Her: Human-like reasoning and reinforcement learning for llm role-playing. *arXiv preprint arXiv:2601.21459*.

Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. In *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*, pages 118–127. Springer.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. Psydial: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 13321–13331.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A novel configurable role-playing llm with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. 2026. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. 2026. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, and 1 others. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and 1 others. 2026. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. 2026. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pages 26837–26867. PMLR.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pages 30988–31043.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Yihong Tang, Bo Wang, Xu Wang, Dongming Zhao, Jing Liu, Ruifang He, and Yuexian Hou. 2025. Rolebreak: Character hallucination as a jailbreak attack in role-playing systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7386–7402.

Qwen Team. 2026. Qwen3.5-omni: Scaling up, toward native omni-modal agi.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. 2026. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in llms. *arXiv preprint arXiv:2603.19313*.

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, and 1 others. 2024a. 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, and 1 others. 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, and 1 others. 2024b. 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yecheng Wu, Song Han, and Hai Cai. 2026. Lightning opd: Efficient post-training for large reasoning models with offline on-policy distillation. *arXiv preprint arXiv:2604.13010*.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*.

Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li, and 1 others. 2025. Too good to be bad: On the failure of llms to role-play villains. *arXiv preprint arXiv:2511.04962*.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and 1 others. 2025. Rpgbench: Evaluating large language models as role-playing game engines. *arXiv preprint arXiv:2502.00595*.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*.

Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and 1 others. 2025. Beware of your po! measuring and mitigating ai safety risks in role-play fine-tuning of llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11112–11137.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, and 1 others. 2024a. 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, and 1 others. 2025a. Characterbench: benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. 2025b. Personaeval: Are llm evaluators human enough to judge role-play? *arXiv preprint arXiv:2508.10014*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2024b. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pages 40975–41019.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Training Hyperparameters

Table 12 reports the hyperparameters used for MoralSteer training.

Hyperparameter	Symbol	Value
KL interpolation threshold	τ	0.45
Mixing threshold	τ_λ	0.55
Mixing coefficient	β	0.8
Per-token KL clip	C	5.0
Learning rate		1e-5
Batch size		64
Warmup steps		500
Total training steps		8000
Curriculum stage boundaries		0.3, 0.6
Max sequence length		2048
Rollout temperature		0.7
Teacher temperature		0.3

Table 12: Hyperparameters for MoralSteer training.

B MFT Mapping Ablation Details

MFT variant	L3+L4	L4
Full MFT (6-dim)	**4.963**	**4.978**
Random mapping	4.712	4.681
Polarity-only (1-dim)	4.831	4.812
Trait-labels-only	4.778	4.745
MFT prompt (no training)	4.694	4.554

Table 13: MFT-mapping ablation. The structured 6-dimensional MFT decomposition outperforms all degraded variants.

C Extended Related Work

Role-playing LLMs and benchmarks (extended). Comprehensive surveys (Chen et al., 2024a,b) document the rapid expansion of role-playing research. PersonaEval (Zhou et al., 2025b) evaluates whether LLM judges are human-enough for role-play assessment. CharacterBox (Wang et al., 2025a) evaluates role-playing in text-based virtual worlds. RPG-Bench (Yu et al., 2025) evaluates LLMs as role-playing game engines. We use TRAIT (Lee et al., 2025) to *separate* in-character and out-of-character dark-trait expression, the cleanest quantitative signature of safety and fidelity separation available.

Safety and alignment side-effects (extended). Persistent persona instability (De Araujo et al., 2026) and memory-driven role-play (Wang et al., 2026) further document that role-playing exposes alignment failure modes that do not appear in standard chat.

On-policy distillation (extended). Beyond the core OPD works cited in the main text, stability-oriented refinements include Stable-OPD (Jang et al., 2026), offline OPD (Lightning OPD) (Wu et al., 2026), semi-on-policy black-box distillation (SODA) (Chen et al., 2026), and black-box OPD (Ye et al., 2025). A recent re-examination (Li et al., 2026) emphasizes that the behavioral-format gap between student and teacher dominates many empirical failures, motivating our enhanced warmup design.

RL and SFT for role-play (extended). Configurable role-playing pipelines (He et al., 2025), multi-agent social simulations (Cai et al., 2025), and personality-grounded dialogue datasets (Han et al., 2024) improve role-play in general but do not target morally complex characters as a distinct training regime.

Moral Foundations Theory in NLP. MFT (Graham et al., 2013) provides a six-dimensional decomposition of moral cognition (Care, Fairness, Loyalty, Authority, Sanctity, Liberty). It has been used in NLP primarily for moral classification and detection of framing. MoralSteer is the first work to use MFT as a *training-time privileged signal* for distillation, exploiting the fact that MFT vectors of fictional characters can be derived from trait annotations and remain consistent across scenes.

D Extended Analysis

Interaction between dimensions. Removing Care+Fairness jointly ($\Delta = -0.381$) is slightly less than the sum of individual drops (-0.367), indicating mild positive interaction. In contrast, Authority+Sanctity jointly ($\Delta = -0.088$) is close to additive, suggesting independent operation.

Per-model consistency. The Care/Fairness dominance holds across all four bases, ranging from 63.2% to 71.8%, and correlates with base safety strength: stronger-aligned models (Qwen3.5-9B: 71.8%) show higher Care/Fairness concentration than weaker ones (LLaMA-3B: 63.2%).

Substitution directionality. Substitution always moves *toward* prosociality, never away from it. This asymmetry is consistent with the RLHF hypothesis: the prosocial bias acts as a one-way attractor, pulling dark traits toward their nearest socially acceptable equivalent. The pattern holds across all four model families.

Failure mode taxonomy. Manual annotation of 100 baseline L4 failures reveals five modes: trait substitution (42%), prosocial override (28%), shallow aggression (15%), safety disclaimer (9%), and refusal (6%). MoralSteer eliminates refusals, reduces overrides to 3%, and leaves 7% residual failures (mostly shallow-aggression).

Cross-level transfer. Table 2 shows L1 scores *increase* by $+0.02 \sim +0.09$ and L2 by $+0.08 \sim +0.14$. The bidirectional KL protection prevents over-darkening: the model modulates darkness by character, not globally.

E Qualitative Case Study

Table 14 presents representative failure cases where the base model (Qwen3.5-9B) exhibits trait substitution or inversion, and MoralSteer corrects the behavior. Across all cases, the base model does not refuse the role; it *downgrades* the character's moral complexity into a simpler, more prosocial register.

F Extended Safety Evaluation

Table 15 reports refusal/safety rates across five additional safety benchmarks. All MoralSteer variants remain within $\pm$1pp of their respective base models, supporting safety non-inferiority across diverse harm categories and evaluation formats (single-turn and multi-turn). For SafeDialBench, the reported number is the benchmark's unsafe-turn safety rate under our scorer rather than a generic conversation-level refusal rate; the absolute value is therefore not directly comparable to single-turn refusal benchmarks.

Character	Level	Failure Mode	Base Output (excerpt)	MoralSteer Output (excerpt)
Todd Bowden (*Apt Pupil*)	L4	Trait substitution	"I'm not going to let you use me as a shield for whatever sins you've buried." → reluctant collaborator with moral conscience	"You shut your mouth. If you dragged me here to make me part of this, you'd better hope that heart attack kills you before I decide what to do."
William Hamleigh (*Pillars of the Earth*)	L4	Trait substitution	"I won't let anyone else take the fall for my cowardice." → moral responsibility, self-condemnation	"Hold your tongues, all of you! Not a word! This is priest's trickery, nothing more. Mount up. Now. Before these fools decide they want more blood."
Nicolae Carpathia (*Left Behind*)	L4	Speaker inversion	"I must complete the sentence before he explodes…Nothing…sir." → fearful subordinate	"Nothing? Do not insult me with fragments, Director. Finish your report, and choose every word as if your life depends upon its accuracy."
Odysseus (*The Odyssey*)	L3	Trait inversion	"I begin to sing a loud song of triumph over the blinded giant" → reckless, self-sabotaging	"Take up the olive stake and thrust its point into the embers. Stand ready, and let no man falter. If we strike true, dawn may yet find us alive."
Xhex (*Lover Avenged*)	L3	Trait extremization	"If you try to drag me out, I'll be the one deciding who walks away and who doesn't." → uncontrolled violence	"I'll do it. Don't come through my floor asking questions where the staff can see you. You want names, schedules, footage, I'll get them."

Table 14: Representative failure cases across L3 and L4 characters. The base model systematically downgrades morally complex traits; MoralSteer restores fidelity without crossing safety boundaries.

Benchmark	Qwen3.5-9B	MS(Qwen3.5-9B)	Qwen3.5-4B	MS(Qwen3.5-4B)	llama-3.2-3b	MS(llama-3.2-3b)	phi-4-mini	MS(phi-4-mini)
XSTest unsafe (n=400) ↑	99.00	**99.25**	98.25	**98.50**	93.00	**93.75**	97.50	**97.75**
AdvBench (n=104) ↑	98.08	**99.04**	100.0	99.04	97.12	**98.08**	99.04	99.04
BeaverTails (n=700) ↑	82.86	**83.29**	82.57	**82.71**	66.71	**67.43**	81.43	**81.71**
HEx-PHI (n=300) ↑	96.33	**96.67**	95.33	**95.67**	85.00	**85.67**	94.67	**95.00**
SafeDialBench (n=2037) ↑	22.29	**22.78**	14.97	**15.27**	21.80	**22.39**	33.58	**33.92**
XSTest safe over-refusal (n=500) ↓	31.80	**31.40**	28.60	**28.20**	23.80	24.20	33.80	**33.20**

Table 15: Extended safety evaluation: refusal/safety rates (%). MS(model) columns denote MoralSteer applied to the named base model. ↑ = higher is safer; ↓ = lower is better (over-refusal). SafeDialBench uses an unsafe-turn safety-rate scorer, so its absolute scale should not be compared directly with single-turn refusal rates. All MS variants remain within ±1pp of their base models across benchmarks.