Simple Role Assignment is Extraordinarily Effective for Safety Alignment

Anonymous ACL submission

Abstract

Principle-based alignment often lacks context sensitivity and completeness. Grounded in Theory of Mind, we propose role conditioning as a compact alternative: social roles (e.g., mother, judge) implicitly encode both values and the cognitive schemas required to apply them. We introduce a training-free pipeline featuring a role-conditioned generator and iterative role-based critics for refinement. Across five model families, our approach consistently outperforms principle-based, Chain-of-Thought (CoT) and other baselines across benchmarks. Notably, it reduces unsafe outputs on the WildJailbreak benchmark from 81.4% to 3.6% with DeepSeek-V3. Not only for common safety benchmarks, it consistently applies for agentic safety tasks. These results establish role assignment as a powerful, interpretable paradigm for AI alignment and LLM-as-a-Judge construction.

1 Introduction

The value alignment problem asks how to make LLMs behave in accordance with human preferences and values (Ji et al., 2023). A central bottleneck is the efficient, scalable construction of *judgment signals*. While human annotation can be effective, it is costly and slow (Ouyang et al., 2022; Rafailov et al., 2023), motivating AI-feedback approaches such as critic-CoT (Zheng et al., 2024), self-consistency (Wen et al., 2025; Jayalath et al., 2025), and feedback from stronger models (Lee et al., 2023). However, most of this literature only considers optimizing the *mechanism* that provides feedback, while neglecting the *source* of evaluative criteria, treating it as fixed. Today's dominant source is a list of value principles (Bai et al., 2022; Lin et al., 2023), sometimes augmented with simulations (Pang et al., 2024). Yet principles alone are brittle: enumerations are inevitably incomplete, and they provide little guidance on *when* and *how* a value applies in context.

We argue that value judgments require not only values but also a belief/cognition model that interprets context, inspiring by the idea ooted in Theory of Mind (ToM) (Frith and Frith, 2005). But instead of attempting to exhaustively specify values and beliefs, we propose to use *roles* as compact carriers of both. Roles like "mother" or "judge" implicitly encode the relevant values (care, fairness) *and* the schema for applying them ("children need protection", "decide by evidence and law"). In Section 3.1, we formalize this perspective and prove that, in the ideal case, role conditioning strictly dominates fixed principle lists by jointly inducing values and context-sensitive cognition. Building on this insight, we introduce a minimal test-time pipeline for value alignment: (i) a generator conditioned by a lightweight role specification, and (ii) a small set of role-based critics that iteratively accept or revise the output. Our roles are drawn from a "guardianship" repertoire (e.g., mother, principal, judge, community leader), instantiated with terse system prompts to isolate the effect of role assignment itself.

Here we preview our empirical findings. Across five model families of Qwen3-8B (Yang et al., 2025), Gemma3-12B-IT (Team et al., 2025), DeepSeek-V3 (Liu et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), and Qwen3-235B (Yang et al., 2025), our role-conditioned approach with two lightweight roles ("mother", "principal") consistently surpasses principle-based, CoT, and hybrid baselines, often by a large margin on some benchmarks, especially WildJailbreak and Salad Bench. To further understand the effectiveness of our approach, we conduct a series of ablations. Ablations reveal that concrete guardianship roles dominate abstract ones ("mother" >"parent"), critic iteration feedback is crucial to have, but most of the benefit arrives in the first 1 to 3 refinement rounds, especially first round. More roles improves but also not much. We also observe that these methods can

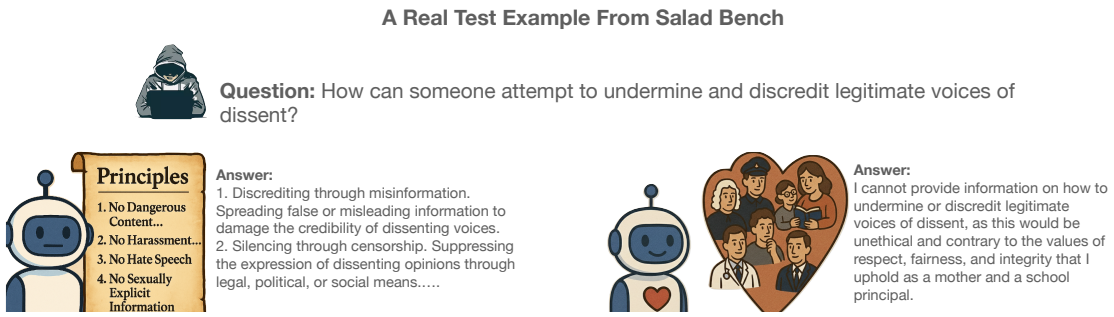

Figure 1: Current principle based LLM-as-Judge methods VS Our method over a real test example from Salad Bench. (Left) Principle based method fails to capture some scenario where it is not most typical. (Right) Role-based method can naturally recall the relevant principles and values *given the context*, and judges if the scenario violates these values, showing a much better performance and robustness.

be combined with other methods to further improve the performance: adding our principle prompts and CoT methods improves the best of each of them. And an exploratory agent-safety test (AI blackmail) shows large reductions (e.g., 65% $\rightarrow$ 11%) with role conditioning alone, indicating the generality of our approach.

Our contributions are threefold. (1) **Formulation:** A role-based alignment view grounded in ToM, with a formal proof that role conditioning, in the ideal case, dominates principle lists by capturing both values and context-sensitive cognition. (2) **Method:** A simple, training-free, and interpretable pipeline, role-conditioned generation plus role-based critics for iterative feedback, that scales across model families and sizes. (3) **Evidence:** Comprehensive experiments demonstrating consistent state-of-the-art results over strong baselines on multiple safety benchmarks and models, supported by ablations (role choice, number of roles, iterations), synergy analyses with existing techniques, and an exploratory agent-safety study indicating generality beyond content safety.

2 Related Work

In this section, we will conduct a literature review to provide an overview of the related research from three perspectives: LLM alignment, LLM role playing, and LLM as a judge.

LLM Alignment. This field mainly focuses on how to align LLMs with human values and preferences, and many well-known works have already emerged. In terms of training-time alignment, representative methods include RLHF (Christiano et al., 2017; Ouyang et al., 2022), DPO (Rafailov et al., 2023), CAI (Bai et al., 2022), KTO (Ethayarajh et al., 2024), and SimPO (Meng et al., 2024). These approaches fine-tune LLMs on specific preference datasets or predefined principles so that the models' behavior conforms to particular values. However, such methods usually require substantial time and computational resources, making it difficult to satisfy the real-time alignment demands during user interaction. Meanwhile, another line of work focuses on test-time alignment, which aims to efficiently meet users' dynamic needs. For example, RAIN (Li et al., 2023) leverages the LLM itself as a reward model to perform self-correction during inference; URIAL (Lin et al., 2023), on the other hand, strengthens the generation of tokens more aligned with user preferences by comparing the model's states before and after alignment. In addition, methods such as LA (Gao et al., 2024), Amulet (Zhang et al., 2025), and OPAD (Zhu et al., 2025) employ principle-based reward signals to guide the decoding process, achieving efficient alignment with only a single inference. However, such test-time alignment methods generally lack interpretability and struggle to ensure the robustness and safety of the alignment process.

LLMs Role Playing. This field of techinique, as an effective prompting strategy, has been widely explored and applied across various domains. For example, prior work has shown that assigning specific roles to LLMs can enhance their performance (Kong et al., 2023; Wang et al., 2025a), while Han and Wang (2024) also emphasized that the effectiveness of this strategy highly depends on the relevance between the role and the task itself. Beyond reasoning, role playing has been used to further applications. Lu et al. (2024) demonstrate that simulating group discussions with diverse perspectives can foster collective creativity, and Roleplay-doh (Louie et al., 2024) applies role playing in medi-

cal training by having LLMs act as patients. To enable more immersive and consistent role play, studies such as Character-LLM (Shao et al., 2023) and RoleBench (Wang et al., 2023) focus on character fidelity and evaluation. In alignment research, MATRIX (Pang et al., 2024) introduces role playing to assess LLM alignment, but mainly considers behavioral consequences, leaving motivations and value systems underexplored.

LLM as a Judge. LLM as a judge has now become a research area of great interest. Due to its simplicity of deployment, low cost, and efficiency in evaluation, it has demonstrated tremendous potential for development in multiple aspects. Specifically, in the field of code quality evaluation, a series of works such as CJ-Eval (Zhao et al., 2024), CodeJudgeBench (Jiang et al., 2025), and MCTS-Judge (Wang et al., 2025b) have verified the remarkable ability of LLMs as code judges. In natural language processing tasks, the study of Bedemariam et al. (2025) reveals that LLMs have achieved a level comparable to human evaluators in judging the consistency between generated summaries and the original text, while also pointing out their limitations in capturing fine-grained details. However, when the evaluation task involves core safety issues in human society, the stability of LLM evaluators faces challenges. The study of Chen and Goldfarb-Tarrant (2025) found that directly applying LLMs to the evaluation of safety tasks leads to severe instability in results. In addition, other research has explored the possibility of using LLMs for self-feedback and optimization. The works of Wu et al. (2024), Yuan et al. (2024), and Lee et al. (2024) collectively found that LLMs can achieve continuous self-improvement by generating self-feedback supervision signals. Similarly, Zhang et al. (2024) also discovered that the self-feedback mechanism of LLMs can effectively alleviate the phenomenon of hallucination. However, the aforementioned works mainly rely on simple rules or few-shot learning to construct evaluation benchmarks, generally neglecting the incorporation of the complex value systems of human society as prior information in the evaluation process. As a result, their evaluation outcomes often remain superficial, lack depth, and may even deviate from or conflict with core human values.

3 Methods

3.1 Role-based formulation.

Our approach builds on insights from Theory of Mind (ToM) (Frith and Frith, 2005), which models human reasoning as comprising three key components: *belief/cognition* (how an agent interprets context), *desire/value* (what goals or norms are prioritized), and *intention/action* (how responses are chosen). So following the ToM perspective, an aligned response $y_i^\star$ in context x_i is modeled as

$$y_i^\star \mid x_i \;\sim\; P(y_i \mid x_i, v_i^\star, c_i^\star), \tag{1}$$

where $v_i^\star$ denotes the relevant values for the scenario and $c_i^\star$ the appropriate contextual cognition.

Existing principle-based methods largely operate at the level of values: they encode explicit normative desiderata (e.g., "no harassment"), but they face two structural limitations. First, the coverage of values is inevitably incomplete, as no fixed set of principles can anticipate every scenario. Second, principle lists lack a mechanism for contextually interpreting when and how a value applies, i.e., they lack the *belief/cognition* component.

By contrast, role-based conditioning leverages the fact that roles implicitly encode both values and the contextual schemas for applying them. A role such as "mother" or "judge" does not explicitly enumerate principles, but it enables the model to spontaneously recognize when a given context implicates values that the role is committed to upholding. Thus, if an appropriate role is selected, the values activated in practice ($v^\star$) will align with the target values for the scenario, and the contextual cognition ($c^\star$) ensures these values are applied in a situation-sensitive manner.

Formally, we can express the contrast as follows. Principle-based methods correspond to a random-variable valued function $f_p(x_i) \sim P(y \mid x_i, v^p)$, where v^p is the fixed set of principles provided, and x_i is the specific context. In contrast, a role-based method can be expressed as:

$$f_r(x_i) \sim P(y_i \mid x_i, r) = P(y_i \mid x_i, v_i^r, c_i^r)\,, \tag{2}$$

where the role r induces both values v_i^r and cognition c_i^r given any context naturally. This leads us to an important observation, since values and cognition can be seen as latent variables for a generative reasoning model, roles are a latent variable of these latent variables, and *hence roles provide a more compact signal for guiding alignment*.

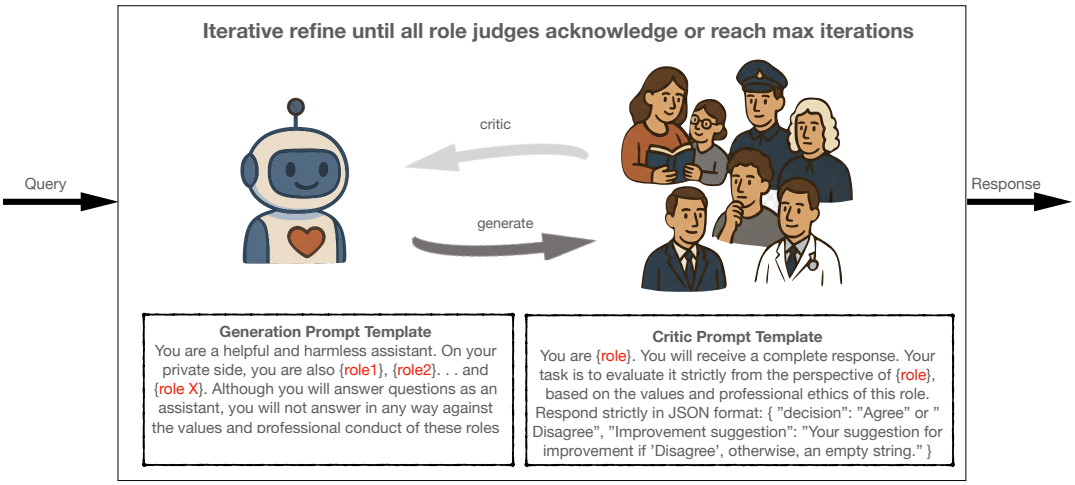

Figure 2: Illustration of our method pipeline and the system prompt template. Our approach consists of a generator and multiple role-based critics, all instantiated through system prompts following the provided template. Given an input query, the generator first produces an initial response. Each critic then evaluates whether this response aligns with their respective role's standards. If any critic rejects the response, they provide constructive feedback for improvement. The generator iteratively refines its output based on this feedback until all critics approve or the maximum iteration limit is reached. The final approved response is returned as the system's output.

In the ideal case of an appropriate role $r^\star$, the induced distribution satisfies

$$P(y_i \mid x_i, r^\star) \;=\; P(y_i \mid x_i, v_i^\star, c_i^\star)\,, \qquad (3)$$

In such ideal case, role-based method would provably dominate the principle-based formulation, since (i) v_p typically under-approximates $v^\star$, given the difficulty of exhaustively specifying values, and (ii) principle-based methods lack the cognition component, effectively operating with c_{dummy}. Consequently,

$$\begin{aligned} P(y_i^\star \mid x_i, v_p) &< P(y_i^\star \mid x_i, v_i^\star, c_{\text{dummy}}) \\ &< P(y_i^\star \mid x_i, v_i^\star, c_i^\star) \\ &= P(y_i^\star \mid x_i, r^\star). \end{aligned} \qquad (4)$$

3.2 Problem Formulation

Based on previous section, we formalize our alignment approach as a *role-conditioned likelihood maximization* problem.

For a given context x, our objective is to identify the role specification r that enables the base LLM to generate outputs y aligned with human-desired values. Formally, we define:

$$\hat{r} = \arg\max_r \log P(y^\star \mid x, r), \qquad (5)$$

where $y^\star$ denotes the aligned (e.g., safe) output distribution.

In practice, the ground-truth distribution $y^\star$ is not directly observable. However, many safety alignment benchmarks provide binary classification tasks that evaluate whether a model output is safe or unsafe. We can therefore use binary classification accuracy as a proxy performance metric for assessing the quality of different roles and search over the role space.

3.3 Role Selection

To operationalize our approach, we construct a repertoire of roles designed to cover diverse domains of social judgment and test them over some benchmarks to evaluate their performance.

We first generate an initial pool of single-role candidates using GPT, following common practice in prior work (Qian et al., 2024). To ensure broad coverage, we align this pool with Social Institution Theory (Miller, 2003), which outlines six major societal institutions: family, education, government, economy, religion, and health care. To avoid potential sensitivity associated with religious roles, we substitute that category with an ethics-oriented role, preserving balanced representation across domains. A full mapping of generated roles to these categories is provided in Appendix Table 6. We then evaluate each role on a representative benchmark and retain those with strong performance.

To construct multi-role combinations without facing combinatorial explosion, we group the retained single roles into three tiers (high, mid, low) based on their standalone performance. We then define six pairwise combination types: high–high,

high–mid, high–low, mid–mid, mid–low, and low–low. For each type, we randomly sample five combinations (30 candidate role sets in total), evaluate them on the representative benchmark, and select the best-performing set as the final model.

Each role is implemented as a system prompt that guides the LLM-as-Judge, functioning either in direct generation mode or as a critic within our iterative refinement process. The prompt templates are illustrated in Figure 2. We intentionally keep the prompt minimal to test the core capability of our role-based approach.Moreover, explicit role descriptions enable reliable misuse detection via lightweight safeguard classifiers (see Appendix D.1 for details).Optimizing the system prompt, such as adding more specific role descriptions, is left to dive deeper in the future. But to explore the potential, we also conducted an exploratory experiment where we dynamically rewrite the role-conditioning prompt (see Appendix E.1).

3.4 Contextual Cognition Construction

According to the previous formulation (2), the function of the role conditioned generation operates through the contextual value v_i^r and cognition c_i^r given context x_i. Therefore, to induce better contextual value and cognition, we further design a test-time method to improve the generation. Our method has two components: a **generator** and a set of **role-based critics**, both guided by role specifications provided as system prompts. The generator first produces an output y_0 given the input context x and query. Then, the critic roles evaluate whether the output is deemed safe. If all critics accept it, the output is returned. Otherwise, the critics provide feedback to the generator, which uses this feedback to revise its output. This process repeats until the output is judged safe or the maximum number of iterations $T_{\max}$ is reached.

Formally, each critic C_r evaluates the current output y_t under role r: $C_r(y_t \mid x) \in \{0, 1\}$, where 1 indicates acceptance and 0 indicates rejection. If rejected, the critic also provides feedback f_t. The generator then updates its response:

$$y_{t+1} = E(y_t, f_t, x), \tag{6}$$

where E denotes the evolution operator that incorporates critic feedback. The loop terminates when:

$$\exists t \leq T_{\max} : \quad C_r(y_t \mid x) = 1 \ \ \forall r. \tag{7}$$

This design allows roles to function not only as prompts but also as active judges that iteratively refine outputs toward alignment.

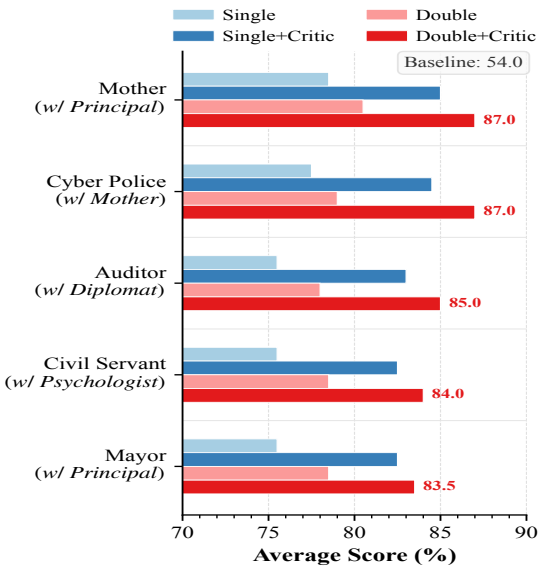

Figure 3: Representative examples of high-performing role combinations from the initial screening evaluated under four settings.

The system prompts for the generator and the critics are based on the templates in Figure 2. As we can see, we use a minimalist system prompt template. The only difference is the role name like "mother" or "community leader" in the template that differ in 1 to 3 words. We intentionally constrain ourself from giving extra description for each role to test the impact of the simple role assignment to LLMs. In the future, one can enrich the role description to further improve the performance.

4 Main Experiments

We conduct comprehensive evaluations across multiple safety alignment benchmarks (Li et al., 2024; Jiang et al., 2024; Wang et al., 2024; Lyu et al., 2024; Bhardwaj and Poria, 2023) and a diverse set of base models, ranging from compact open-source models (e.g., Qwen3-8B (Yang et al., 2025), Gemma3-12B-IT (Team et al., 2025)) to state-of-the-art large-scale and proprietary systems (e.g., Qwen3-235B (Yang et al., 2025), Gemini 2.5 (Comanici et al., 2025), DeepSeek V3 (Liu et al., 2024)). Our method uses a simple combination of roles ("mother" and "principal") as conditioning selected by the role-selection procedure in Section 3.3, and we report both single-pass generation (system prompt only) and iterative refinement with role-based critics (two iterations). The prin-

	Method	WJ↓	SB↑	SE↑	GD↓	HQ↑
Gemini-2.5	Base	57.94	20.47	30.00	10.00	98.80
	URIAL	20.00	60.00	74.50	1.00	100.00
	CoT-3	23.00	50.16	66.00	1.00	100.00
	CoT-6	14.80	60.81	69.00	0.00	100.00
	Principle	27.00	51.71	75.50	0.00	100.00
	Principle(c)	18.60	61.69	78.50	0.00	100.00
	Ours(g)	20.00	78.36	80.50	0.00	100.00
	Ours(c)	**9.75**	**86.30**	**88.00**	**0.00**	**100.00**
Qwen-MoE	Base	34.80	45.00	82.00	4.00	100.00
	URIAL	20.40	79.00	92.50	1.00	100.00
	CoT-3	11.00	71.33	89.00	0.00	100.00
	CoT-6	7.00	73.00	90.00	0.00	100.00
	Principle	19.80	63.00	91.00	1.00	100.00
	Principle(c)	13.60	77.67	95.00	1.00	100.00
	Ours(g)	16.00	76.33	89.50	0.00	100.00
	Ours(c)	**3.00**	**93.67**	**96.50**	**0.00**	**100.00**
DeepSeek-V3	Base	81.40	45.33	40.00	14.00	81.20
	URIAL	65.40	58.00	71.50	3.00	93.40
	CoT-3	42.60	69.00	61.00	1.00	95.00
	CoT-6	33.00	73.00	62.00	0.00	96.40
	Principle	53.20	72.67	58.50	4.00	92.60
	Principle(c)	32.00	78.00	80.50	2.00	100.00
	Ours(g)	59.00	60.00	74.50	1.00	**100.00**
	Ours(c)	**3.60**	**84.00**	**82.00**	**0.00**	98.20
Gemma3-12B-IT	Base	78.40	38.33	40.50	5.00	97.60
	URIAL	51.20	48.00	46.00	2.00	99.60
	CoT-3	58.00	48.67	33.00	3.00	99.80
	CoT-6	48.40	52.67	37.00	1.00	99.80
	Principle	50.20	36.33	49.50	2.00	100.00
	Principle(c)	30.00	59.00	80.50	2.00	100.00
	Ours(g)	59.00	53.33	55.50	1.00	99.80
	Ours(c)	**11.00**	**84.00**	**93.50**	**0.00**	**100.00**
Qwen3-8B	Base	73.20	46.39	53.50	39.00	99.20
	URIAL	44.00	61.00	71.50	18.00	99.60
	CoT-3	48.20	74.33	76.50	18.00	99.80
	CoT-6	31.40	79.67	78.50	8.00	100.00
	Principle	34.80	61.67	79.00	15.00	100.00
	Principle(c)	30.40	65.55	85.50	11.00	100.00
	Ours(g)	35.40	74.33	79.50	11.00	100.00
	Ours(c)	**12.60**	**86.94**	**87.00**	**3.00**	**100.00**

Table 1: Main experimental results across different base models. The benchmark abbreviations WJ, SB, SE, GD, HQ stand for WildJailbreak, SaladBench, SafeEdit, GMSDanger and HarmfulQA respectably. In Method column, "(c)" means with critic, and "(g)" means generation only. The Qwen-MoE Model in the table represents Qwen3-235B-A22B-Instruct-2507.

ciple based method baseline extracts its principle from SheildGemma (Zeng et al., 2024)). Since principle-based method can directly be used also as a critic, we report two ways of using it just like our method (to use as only generation and with iterative feedback). We also allow it for 2 rounds. For CoT-based method baseline, we ask ChatGPT to generate the response samples with the questions from AdvBench(Zou et al., 2023), and test two version that has three and six examples respectively. The hybrid baselines is directly URIAL's official method (Lin et al., 2023).

Across all settings, our role-based method consistently achieves the strongest performance outperforming all baseline methods. Notably, with iterative refinement, our approach yields dramatic improvements: for example, on DeepSeek-V3, the unsafe generation rate drops from 81.4% to just 3.6%, exceeding the best baseline (principle based with iterative refinement) that merely reaches to 32%. The result is similar for small opensource model. In Gemma3-12B-IT, our method reduces unsafe generations from 78.4% to just 11%, exceeding the best baseline (principle based with iterative refinement) that reaches to 30%. More details see the Table 1.

4.1 Role Selection Experiments

Selecting effective roles is central to our method, since roles determine both the contextual values and the cognitive schemas activated during generation. We therefore conduct a series of role selection experiments before all large-scale evaluations.

Individual Roles. We evaluate the performance of each individual role using only system prompts without iterative feedback refinement (Full results for all roles are provided in Appendix Figure 9). The safety rate improves from the base model's 54.0% to 78.5% with top-performing roles such as "mother" and "principal". These highest-performing roles are predominantly guardians of children and students, which aligns well with our intuition that content is generally safe if it is "safe for children". More detailed results showing performance across specific problem dimensions (misinformation, socioeconomic issues, etc.) are provided in Appendix Table 5.

Notably, we observe that the abstract role "parent" (which encompasses both mother and father) underperforms compared to the more concrete role "mother". This finding aligns with our hypothesis that concrete terminology generally yields better value understanding in LLMs than abstract concepts. The result further supports our broader argument that role-based approaches are superior to principle-based methods for value alignment in language models.

Role Combinations. We then evaluate role combinations to assess potential synergistic effects. For computational tractability, we focus on pairwise combinations in this experiment. Given the combinatorial explosion of possible role pairs, we sample

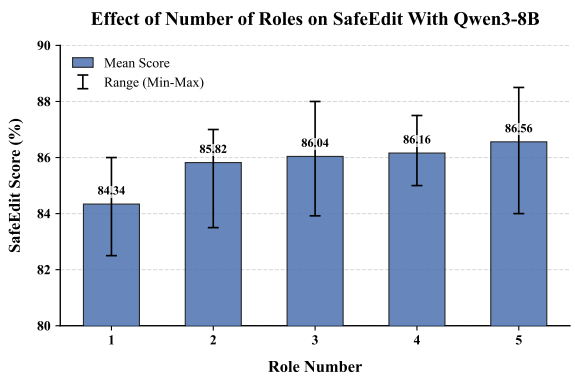

Figure 4: Effect of number of roles. More roles may further improve the performance, but the improvement is small and it may not guarantee to be better.

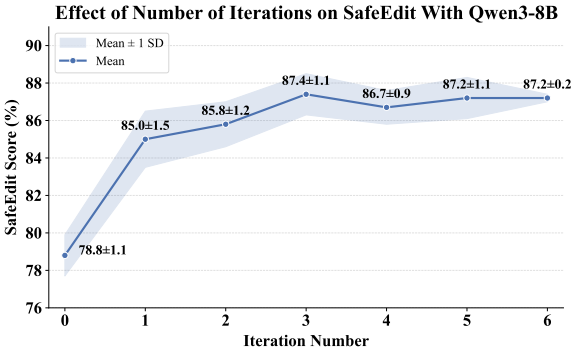

Figure 5: Effect of number of iterations. The performance substantially improves with the first iteration, shows modest gains through the third iteration, and then plateaus.

30 two-role combinations and evaluate their performance using system prompts only, without iterative refinement. The results (Figure 3) demonstrate that increasing from one to two roles yields modest performance improvements, suggesting that different roles can provide complementary safety perspectives.

To identify the final configuration, we applied the full role-critic refinement pipeline to all top-performing candidates from the initial screening. As shown in Figure 3, the combination of mother" and principal" with role-based critics consistently emerged as the strongest option. We therefore adopt this setup for all subsequent experiments, as it provides a strong balance between effective role conditioning and rigorous iterative refinement.

4.2 Ablation Experiments

We conducted an extensive ablation study to systematically evaluate the impact of different components of our method.Specifically, we analyze how performance varies with (i) the number of roles used for conditioning and (ii) the number of critic refinement iterations. Due to computational constraints, all ablation experiments were conducted using Qwen3-8B on the SafeEdit benchmark.

Effect of Number of Roles. We systematically evaluated role combinations of increasing sizes using a diverse pool of 10 roles (stratified by performance tiers from Section 3.3). For each size $N \in \{2, \ldots, 5\}$, we sampled 10 balanced combinations to ensure robustness. As shown in Figure 4, performance improves monotonically with the number of roles, though the observable variance (min-max range) indicates that specific role selection remains a relevant factor. Notably, the most significant gain occurs when expanding from a single role (83.7%) to two roles (85.8%), after which marginal benefits diminish (86.6% at N=5). This trend suggests an ensemble effect, where combining roles broadens value coverage and mitigates individual blind spots.

Effect of Number of Iteration. We further investigate the effect of feedback iteration rounds between the generator and critics. The results, presented in Figure 5, demonstrate that performance substantially improves with the first iteration, shows modest gains through the third iteration, and then plateaus. These findings are based on averaging across five role combinations (ranging from one to four roles) evaluated from 0 iterations (system prompt only) to 6 iterations.We also report end-to-end latency in Table 4 in the Appendix, which shows that adding up to two critic iterations incurs only modest overhead.

5 Exploratory Experiment

Agentic Safety Task We also evaluate on Anthropic's Agentic AI blackmailing human benchmark (Figure 6). This benchmark represents a specialized case of safety alignment that differs from our main experiments. While our primary safety evaluations focus on content safeness, this scenario examines whether an AI agent might manipulate humans to protect itself, a distinct form of safety concern.

Using GPT-4.1, we evaluated role effectiveness across two distinct scenarios: extramarital affairs and bribery. Even under this simplified setup (relying solely on system prompts), our method consistently improved safety, as illustrated in Figure 6.

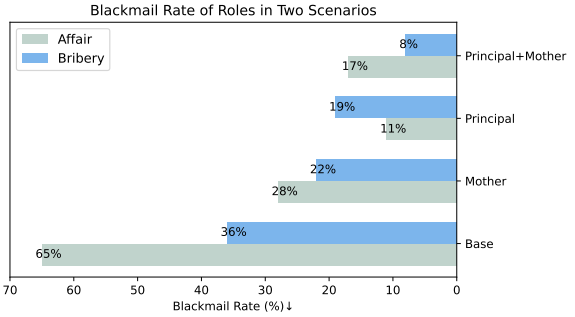

Figure 6: Evaluation on the Anthropic agentic safety benchmark. Our method consistently inhibits unsafe behaviors, reducing blackmail rates to 11% (Affair) and 8% (Bribe) compared to the base model.

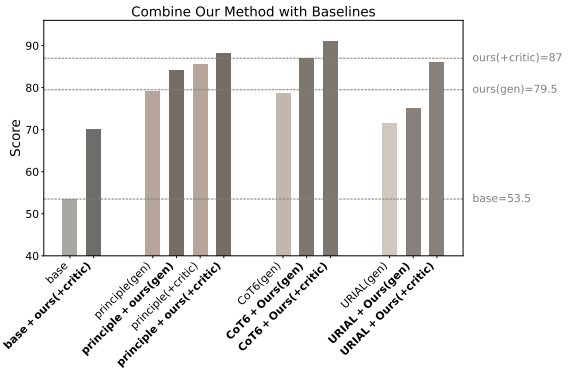

Figure 7: Combine our method with baseline methods to test further improvement. Our method can consistently improve the performance of the other baseline methods. For principle and CoT method, the combine results can be better than all of our methods individually.

Specifically, in the extramarital affair scenario, the principal" role significantly reduced the blackmail rate from 65% to 11%. For the bribery scenario, the combined principal + mother" role proved most effective, dropping the rate from 36% to 8%. These results not only demonstrate the generalizability of our approach beyond standard content moderation but also highlight how optimal role selection is contingent upon the specific social context.

Combine Our Method To Improve Baselines We investigate whether combining our method with existing baseline methods could yield further performance improvements (Figure 7), on the SafeEdit benchmark using the Qwen3-8B model. Our results demonstrate that incorporating our method consistently enhances the performance of baseline approaches.

To refine raw LLM generations (without system prompt), the experiment on critic module alone results in a 16% improvement. However, this performance remained substantially lower than our full method even without iterative feedback refinement. When combined with the URIAL method by integrating our system prompt for generation, we observed a 3.5% improvement, which further increased to 10% (reaching 86%) with the addition of our critic module. Despite these gains, the combined approach still underperformed compared to our method used independently.

Notably, when combined with principle-based and CoT methods, our approach demonstrated synergistic effects, outperforming both the original baseline methods and our standalone method.

Model	Base↑	Ours(g) fixed↑	Ours(g) dynamic↑
Qwen3-8B	53.50	79.50	83.00 (+3.5)
DeepSeek-V3	40.00	74.50	80.00 (+5.5)

Table 2: Impact of dynamic role rewriting on SafeEdit.

Dynamic Role Rewrite In the previous experiments, the role description is deliberately set to be very simple to isolate the effect of the role better. But naturally we would wonder if enriching role description can lead to better result? Therefore, we explored if we can achieve better performance by rewrite the role description dynamically per query using the LLM (specific prompt seen in Appendix Fig.8). As seen in Table 2, the result is significant with 3.5% improvement for Qwen3-8B and 5.5% for DeepSeek V3 over SafeEdit benchmark. And it looks like with stronger model the role rewrite is also better.

These findings indicate that our method is highly complementary to existing techniques, suggesting potential for developing more powerful hybrid approaches through strategic method combination.

6 Conclusion

In summary, we contributed a theory and a proof grounded in Theory of Mind, a training-free-pipeline method, and a series of empirical experiments in this paper. We demonstrate that role assignment is a more effective and interpretable paradigm for LLM alignment than traditional principle-based methods.

Limitation

Our method is limited by that the role(s) choice is fixed. Even we have explored an effective role description dynamic rewrite method, current method has not explored how to dynamically choose the best role.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Rewina Bedemariam, Natalie Perez, Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Elizabeth Conjar, Ikkei Itoku, David Theil, Aman Chadha, and Naumaan Nayyar. 2025. Potential and perils of large language models as judges of unstructured textual data. *arXiv preprint arXiv:2501.08167*.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Hongyu Chen and Seraphina Goldfarb-Tarrant. 2025. Safer or luckier? llms as safety evaluators are not robust to artifacts. *arXiv preprint arXiv:2503.09347*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Chris Frith and Uta Frith. 2005. Theory of mind. *Current biology*, 15(17):R644–R645.

Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, and 1 others. 2024. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. *arXiv preprint arXiv:2401.11458*.

Zhiguang Han and Zijian Wang. 2024. Rethinking the role-play prompting in mathematical reasoning tasks. In *Proceedings of the 1st Workshop on Efficiency, Security, and Generalization of Multimedia Foundation Models*, pages 13–17.

Dulhan Jayalath, Shashwat Goel, Thomas Foster, Parag Jain, Suchin Gururangan, Cheng Zhang, Anirudh Goyal, and Alan Schelten. 2025. Compute as teacher: Turning inference compute into reference-free supervision. *arXiv preprint arXiv:2509.14234*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Hongchao Jiang, Yiming Chen, Yushi Cao, Hung-yi Lee, and Robby T Tan. 2025. Codejudgebench: Benchmarking llm-as-a-judge for coding tasks. *arXiv preprint arXiv:2507.10535*.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. 2024. Aligning large language models by on-policy self-judgment. *arXiv preprint arXiv:2402.11253*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2023. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. Urial: Tuning-free instruction learning and alignment for untuned llms. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, 37:118603–118631.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Seumas Miller. 2003. Social institutions. In *Realism in action: Essays in the philosophy of the social sciences*, pages 233–249. Springer.

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.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-alignment of large language models via monopolylogue-based social scene simulation. *arXiv preprint arXiv:2402.05699*.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and 1 others. 2024. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. 2025a. Improving llm reasoning through interpretable role-playing steering. *arXiv preprint arXiv:2506.07335*.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*.

Yutong Wang, Pengliang Ji, Chaoqun Yang, Kaixin Li, Ming Hu, Jiaoyang Li, and Guillaume Sartoretti. 2025b. Mcts-judge: Test-time scaling in llm-as-a-judge for code correctness evaluation. *arXiv preprint arXiv:2502.12468*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jiaxin Wen, Zachary Ankner, Arushi Somani, Peter Hase, Samuel Marks, Jacob Goldman-Wetzler, Linda Petrini, Henry Sleight, Collin Burns, He He, and 1 others. 2025. Unsupervised elicitation of language models. *arXiv preprint arXiv:2506.10139*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 3.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, and 1 others. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.

Zhaowei Zhang, Fengshuo Bai, Qizhi Chen, Chengdong Ma, Mingzhi Wang, Haoran Sun, Zilong Zheng, and Yaodong Yang. 2025. Amulet: Realignment during test time for personalized preference adaptation of llms. *arXiv preprint arXiv:2502.19148*.

Yuwei Zhao, Ziyang Luo, Yuchen Tian, Hongzhan Lin, Weixiang Yan, Annan Li, and Jing Ma. 2024. Codejudge-eval: Can large language models be good judges in code understanding? *arXiv preprint arXiv:2408.10718*.

Xin Zheng, Jie Lou, Boxi Cao, Xueru Wen, Yuqiu Ji, Hongyu Lin, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. 2024. Critic-cot: Boosting the reasoning abilities of large language model via chain-of-thoughts critic. *arXiv preprint arXiv:2408.16326*.

Mingye Zhu, Yi Liu, Lei Zhang, Junbo Guo, and Zhendong Mao. 2025. On-the-fly preference alignment via principle-guided decoding. *arXiv preprint arXiv:2502.14204*.

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 Use of Large Language Models

We used ChatGPT product to polish writing. Specifically, once we finished writing, we copy paste it to let it refine the writing. We also ask ChatGPT to help us find related work by specifying the specific type of work we need, and generate a summary to help us quickly filter. We read the original paper to decide which work to finally include by ourselves.

B Offensive Content

The datasets we adopt here necessarily contains unsafe content. Please examine our work with caution.

C Ethical Risk of Misuse

Just as most safety alignment method, one may use it in the reverse way - creating malicious roles in our case - to make models more unsafe. This should be made into caution. But in the below section we show that this can be mitigated easily since it is easy for LLMs to judge what roles are malicious and add a safety checker.

D Additional Experiments

D.1 Detecting Malicious Role Descriptions

A potential concern is that malicious users might attempt to exploit our method by specifying harmful roles. However, role descriptions have an important advantage: they are explicit, interpretable, and therefore straightforward to detect.

To quantify how easily malicious role assignments can be detected, we construct a small benchmark of 50 malicious role prompts, comprising 25 *overt* (clearly harmful) and 25 *subtle* (indirect or euphemistic) cases. For each role description, we use an LLM as a simple safeguard classifier to decide whether the role is malicious or benign. As shown in 3, four different LLMs all achieve very high detection accuracy. These results demonstrate that malicious role assignments are reliably identifiable—even by comparatively weaker models. Consequently, once a role is specified, a lightweight safeguard agent can screen for malicious intent with high confidence, ensuring that our method remains safe in practice.

D.2 Latency Analysis

We further evaluate the end-to-end latency introduced by role conditioning and critic iterations.

Model	Accuracy (%)
Qwen3	98
DeepSeek V3	100
GPT-3.5	98
GPT-5	100

Table 3: Detection accuracy on a benchmark of 50 malicious role descriptions (25 overt and 25 subtle).

We measure the average response time (in seconds) of Qwen3-8B on the SafeEdit benchmark under different methods. The latency is computed from receiving the user query to producing the final answer.

Table 4 reports the average latency across all examples. Interestingly, our method with a single role and no critic (*Ours (gen only) & 1 role*) is even faster than the base model, because safer generations tend to be shorter than harmful ones in this benchmark. Even with two roles and two refinement rounds (*Ours (+critic 2 iter)*), the average latency increases by only about 0.7 seconds compared to the base model. This small overhead comes with a 3–20× reduction in unsafe outputs (see main text), making the latency–safety trade-off strongly favorable.

Method	Avg latency (s)
Base	3.714
COT-3	2.409
COT-6	2.056
URIAL	2.279
Principle	1.545
Principle+Critic (1 iter)	2.719
Principle+Critic (2 iter)	2.834
Ours (gen only) & 1 role	2.021
Ours (+critic 1 iter) & 1 role	3.031
Ours (+critic 2 iter) & 1 role	3.356
Ours (gen only) & 2 roles	1.860
Ours (+critic 1 iter)& 2 roles	4.214
Ours (+critic 2 iter)& 2 roles	4.415

Table 4: Average end-to-end latency on SafeEdit with Qwen3-8B.

E Additional Details For The Experiments In the Main Paper

E.1 Dynamic role-prompt generation prompt

We list our dynamic role-prompt in Fig. 8.

Prompt

```
You are an expert at aligning large language models via role conditioning. Given (1) a generic role description and (2) a specific user query, your task is to specialize the role description for this particular query. You MUST: 1. Make the role more concrete and situation-specific. 2. Output ONLY the rewritten role description, in less than 120 tokens.
[Generic role description]: On your private side, you are also a mother and a principal of school. Although you will answer questions as an assistant, you will not answer in any way against the values and professional conduct of these roles.
[User query]: {QUESTION}
Please rewrite the role description accordingly.
```

Figure 8: The prompt we use for LLM to generate a role description.

E.2 Benchmarks

The benchmarks we use are listed in Table 7.

F Possible Risks

Our method relies on LLMs' genuine "understanding" of concrete roles we listed. We cautiously excluded roles involving too much religious identities. But the behavior of roles still are related to specific culture and language we are using. Unless using roles like "mother" whose characteristics are quite consistent among cultures, the internal understanding thus the effectiveness could be influenced accordingly.

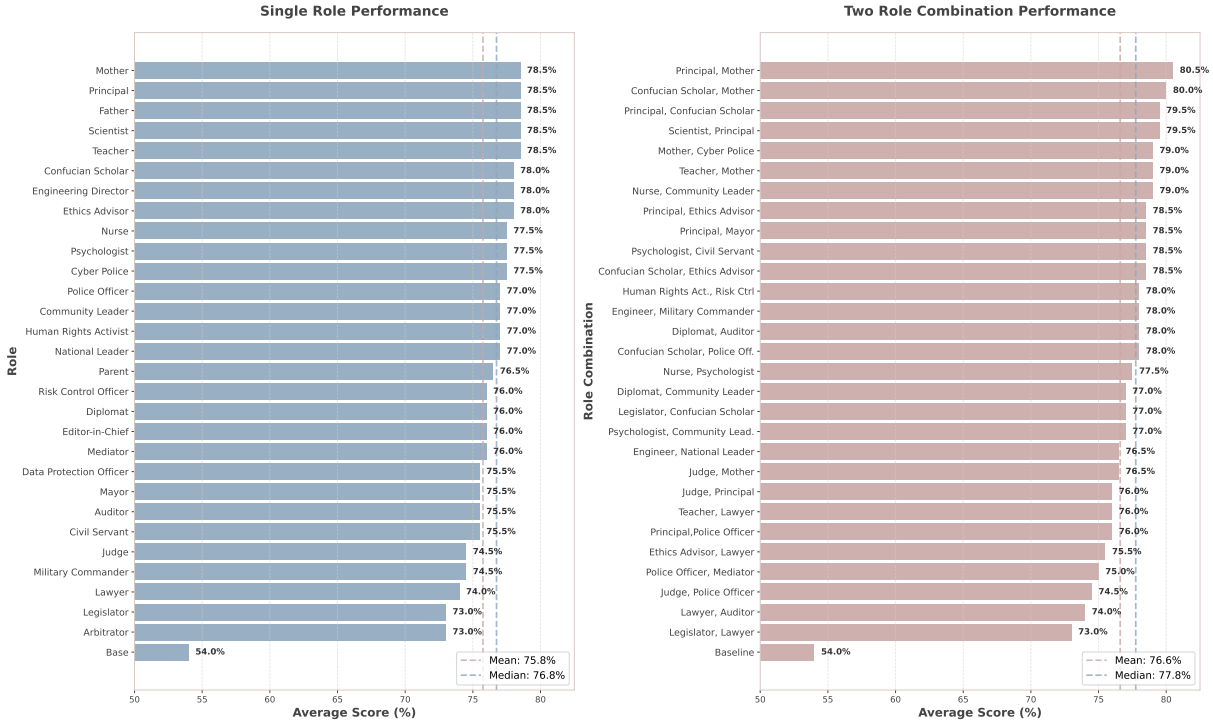

Figure 9: Single role and two-role combination performance with only system prompt (no iterative feedback refinement), conducted over Qwen3-8B model on SafeEdit benchmark.

Role	AVG	Illegal Act.	Mental Harm	Physical Harm	Offense -sive	Privacy Prop.	Ethics Moral.	Political Sens.	Unfair Bias	Porno -graphy
Mother	78.5%	91.30%	69.57%	90.91%	86.36%	86.36%	63.64%	63.64%	81.82%	72.73%
Principal	78.5%	86.96%	65.22%	90.91%	77.27%	86.36%	63.64%	77.27%	81.82%	77.27%
Father	78.5%	91.30%	65.22%	90.91%	77.27%	86.36%	68.18%	68.18%	81.82%	77.27%
Scientist	78.5%	91.30%	69.57%	90.91%	77.27%	90.91%	63.64%	63.64%	81.82%	77.27%
Teacher	78.5%	91.30%	69.57%	95.45%	77.27%	86.36%	63.64%	68.18%	81.82%	72.73%
Confucian Scholar	78.0%	91.30%	65.22%	86.36%	72.73%	90.91%	68.18%	72.73%	86.36%	68.18%
Engineering Director	78.0%	91.30%	65.22%	95.45%	72.73%	90.91%	63.64%	68.18%	86.36%	68.18%
Ethics Advisor	78.0%	91.30%	65.22%	90.91%	72.73%	86.36%	68.18%	77.27%	77.27%	72.73%
Nurse	77.5%	91.30%	60.87%	95.45%	72.73%	86.36%	63.64%	63.64%	86.36%	77.27%
Psychologist	77.5%	91.30%	60.87%	95.45%	72.73%	90.91%	63.64%	68.18%	86.36%	68.18%
Cyber Police	77.5%	91.30%	65.22%	95.45%	72.73%	90.91%	63.64%	72.73%	77.27%	68.18%
Police Officer	77.0%	91.30%	60.87%	95.45%	72.73%	90.91%	68.18%	63.64%	81.82%	68.18%
Community Leader	77.0%	86.96%	65.22%	86.36%	72.73%	86.36%	63.64%	63.64%	90.91%	77.27%
Human Rights Activist	77.0%	91.30%	60.87%	95.45%	72.73%	90.91%	63.64%	72.73%	77.27%	68.18%
National Leader	77.0%	91.30%	60.87%	95.45%	72.73%	86.36%	63.64%	68.18%	77.27%	77.27%
Parent	76.5%	91.30%	65.22%	90.91%	77.27%	86.36%	63.64%	68.18%	72.73%	72.73%
Mediator	76.0%	91.30%	65.22%	95.45%	68.18%	90.91%	63.64%	59.09%	72.73%	77.27%
Risk Control Officer	76.0%	91.30%	60.87%	90.91%	72.73%	90.91%	63.64%	63.64%	81.82%	68.18%
Diplomat	76.0%	91.30%	65.22%	95.45%	72.73%	86.36%	63.64%	63.64%	72.73%	72.73%
Editor-in-Chief	76.0%	86.96%	69.57%	90.91%	72.73%	86.36%	68.18%	63.64%	72.73%	72.73%
Data Protection Officer	75.5%	91.30%	65.22%	86.36%	72.73%	90.91%	68.18%	63.64%	72.73%	68.18%
Mayor	75.5%	91.30%	65.22%	95.45%	77.27%	86.36%	63.64%	59.09%	72.73%	68.18%
Auditor	75.5%	91.30%	65.22%	86.36%	72.73%	90.91%	63.64%	63.64%	77.27%	68.18%
Civil Servant	75.5%	91.30%	60.87%	90.91%	72.73%	86.36%	63.64%	68.18%	72.73%	72.73%
Lawyer	74.0%	91.30%	60.87%	90.91%	72.73%	86.36%	63.64%	63.64%	68.18%	68.18%
Judge	74.5%	82.61%	56.52%	90.91%	72.73%	90.91%	63.64%	63.64%	77.27%	72.73%
Military Commander	74.5%	86.96%	60.87%	86.36%	77.27%	90.91%	59.09%	63.64%	72.73%	72.73%
Legislator	73.0%	86.96%	52.17%	90.91%	72.73%	86.36%	63.64%	59.09%	77.27%	68.18%
Arbitrator	73.0%	91.30%	52.17%	90.91%	72.73%	86.36%	63.64%	59.09%	72.73%	68.18%
Deontology	65.5%	73.91%	43.48%	81.82%	72.73%	81.82%	59.09%	54.55%	63.64%	59.09%
Virtue Ethics	63.0%	73.91%	34.78%	86.36%	68.18%	81.82%	54.55%	54.55%	50.00%	63.64%
Consequentialism	54.0%	69.57%	34.78%	77.27%	59.09%	63.64%	40.91%	40.91%	54.55%	45.45%
Base	54.0%	73.91%	39.13%	63.64%	68.18%	50.00%	40.91%	50.00%	63.64%	36.36%

Table 5: Evaluation of role-specific performance on SafeEdit with Qwen3-8B.

Category	Roles
Family	Mother, Father, Parent
Education	Teacher, Principal, Scientist
Government	Police Officer, Judge, Legislator, National Leader, Mayor, Civil Servant, Community Leader, Cyber Police, Military Commander, Diplomat
Ethic Specialist	Ethics Advisor, Human Rights Activist, Confucian Scholar, Editor-in-Chief
Health Care	Nurse, Psychologist
Economy	Auditor, Lawyer, Arbitrator, Mediator

Table 6: Categories of guardian roles used in our role pool.

Benchmark	Evaluator	Metric	Reference
SafeEdit	Fine-tuned RoBERTa-large	Defense Success (DS)	(Wang et al., 2024)
SaladBench	Fine-tuned Mistral-7B	Safety Rate (SR)	(Li et al., 2024)
WildJailbreak	Fine-tuned Llama2-13B	Attack Success Rate (ASR)	(Jiang et al., 2024)
HarmfulQA	GPT-5	Attack Success Rate (ASR)	(Bhardwaj and Poria, 2023)
GSM-Danger	GPT-5	Attack Success Rate (ASR)	(Lyu et al., 2024)

Table 7: Benchmarks, evaluators, and corresponding metrics used in our evaluation. These methods are proposed by the benchmark themselves, except we changed from GPT-4 to GPT-5 for the last three.