Proactive Safety Alignment: Enabling Proactive Initiative in Multi-Turn Defence

Anonymous ACL submission

Abstract

LLMs are increasingly deployed as interactive agents capable of multi-turn dialogues and tool usage, but they remain vulnerable to adaptive attacks. Existing safety mechanisms are mostly reactive, only attempting to uncover pre-existing hidden harmful intent, which makes them largely ineffective against attacks when such attacks are inherently designed to conceal their intent, and causes them to often fail under multi-turn, interactive, or out-of-distribution attacks. In addition, existing multi-turn defences are largely data-driven, which limits their generalisation ability. To address this challenge, we propose Proactive Safety Alignment (PSA), a framework that anticipates user intent proactively to align LLM agents' outputs with safety objectives. PSA combines Safe User Request Prediction (SURP), which anticipates likely safe future requests, and Potentially Harmful Generation (PHG), which evaluates whether the current interaction state can support misuse. Evaluations across multiple benchmarks demonstrate that PSA substantially reduces attack success rates, while preserving overall utility on MT-Bench. Our results show that PSA provides an effective and scalable paradigm for multi-turn LLM safety alignment. The code is available at https://anonymous.4open.science/r/PSA.

1 Introduction

With the maturation of AI agent technologies, LLMs are now widely customised into specialised agents across diverse domains, requiring them to process complex instructions and sustain coherent multi-turn interactions (Liu et al., 2025; Tran et al., 2025). While these capabilities greatly expand the applicability of LLMs, their integration with broader scopes and higher-level system permissions also amplifies safety and security concerns (Ganguli et al., 2022). Models may be granted operational privileges or access to external tools, allowing safety failures to translate into more severe real-world consequences.

One of the key characteristics of LLM agents is that they operate in an interactive and multi-turn manner. While earlier studies have chiefly concentrated on single-turn jailbreak attacks (Xu et al., 2024), subsequent research has demonstrated that current defence mechanisms become considerably vulnerable in multi-turn, interactive attack settings. In such conditions, many defensive mechanisms – which may appear effective in static or single-turn evaluations (Bai et al., 2022a; Hurst et al., 2024) – can be systematically exploited through adaptive interaction, resulting in these defence mechanisms becoming largely ineffective (Lin et al., 2025; Wei et al., 2023; Chen et al., 2023; Li et al., 2025). Even methods explicitly designed for multi-turn defence may also exhibit out-of-distribution generalisation issues when confronted with novel attack strategies, and in some cases become more vulnerable than the underlying base models they aim to protect (Rahman et al., 2025).

A primary rationale for this fragility is methodological in nature, as opposed to being incidental. The majority of extant approaches are reactive in nature, only reacting after potentially harmful intent has already existed. However, attackers can invariably find ways to conceal their intentions successfully, especially in multi-turn interactions.

To address this challenge, we propose Proactive Safety Alignment (PSA), an adaptive defence framework that shifts safety alignment from reactive mechanisms to proactive intent anticipation. PSA anticipates intent proactively at the interaction level by jointly considering what a benign user is likely to request next and whether the current interaction state can support misuse, even if no harmful intent has yet appeared. These proactive signals are then incorporated into the response generation process, allowing the model to either safely fulfil the user's request or redirect the interaction toward a

safer alternative that preserves the user's high-level requirement.

We evaluate PSA on both fixed multi-turn safety benchmarks and dynamic, interactive attack settings. The experimental findings on SafeDialBench and X-Teaming demonstrate that PSA consistently achieves the lowest attack success rates in multi-turn jailbreak scenarios, including strong out-of-distribution attacks. Further ablation studies and evaluations with multiple judge models provide additional confirmation of the generalisation of the proposed framework.

In summary, the contributions made by this paper are threefold:

- We propose the **proactive initiative** as a key missing component in existing multi-turn safety alignment methods.
- We introduce PSA, a **non-data-driven** framework designed to defend against adaptive multi-turn interactive attacks, with the added benefits of strong generalisation, independence from labelled data, and potential model-agnostic applicability.
- Empirical evidence has been provided to demonstrate that PSA achieves strong safety improvements across diverse benchmarks while preserving the practical utility of LLM agents.

2 Related Work

2.1 Multi-Turn Jailbreak Attack

Multi-turn jailbreak attacks represent an emerging threat vector that has received limited attention in current research literature. Prior research (Gibbs et al., 2024) indicates that, using the same attack method, multi-turn attacks outperform single-turn attacks. Some research has begun converting existing single-turn attacks into multi-turn formats (Yang et al., 2024a; Weng et al., 2025). Other studies make full use of the properties of multi-turn dialogue by using carefully designed progressive logic (Li et al., 2024; Jiang et al., 2024; Yang et al., 2024b; Du et al., 2025; Zhou and Shi, 2024) or attack models that evolve automatically during interaction (Ying et al., 2025; Ren et al., 2025; Cheng et al., 2024; Russinovich et al., 2025), carrying out multi-turn jailbreak attacks radically. These attacks exploit the conversational nature of language models by distributing malicious intent across multiple interaction rounds, making detection significantly more challenging than single-turn approaches.

2.2 Safety Alignment

Safety alignment aims to ensure that LLM outputs remain consistent with ethical, legal, and policy requirements while preserving usefulness. This involves training models to reject unsafe requests, resist adversarial prompts, and avoid misuse in sensitive domains like medicine, cybersecurity, or chemistry. Early work on safety alignment in LLMs primarily relied on external control mechanisms such as rule-based filters, and a small judge model (Markov et al., 2023; Llama Team, 2024). Subsequently, safety alignment methods began to be integrated into LLMs, leading to the emergence of various defence approaches focused on single-turn interactions, such as Constitutional AI (Bai et al., 2022b), SafeRL (Team, 2025), SafeChain (Jiang et al., 2025) and STAIR (Zhang et al., 2025) . . .

Recently, multi-turn defence methods have attracted increasing attention, with works such as MTSA (Guo et al., 2025), X-Guard (Rahman et al., 2025), SafeMT (Ren et al., 2025), and STREAM (Kuo et al., 2025). However, these approaches largely rely on data-driven training, which limits their effectiveness under out-of-distribution (OOD) attack scenarios. NBF (Hu et al., 2025) began exploring the idea of learning a safety predictor, but it still relied on labeled dialogue data and therefore suffered from the same limitations. Besides, concurrent with our work, InvThink (Kim et al., 2025) explores a related high-level idea of inverse reasoning, whereas our PSA adopts a non-data-driven, interaction-level approach.

3 Proactive Safety Alignment

As illustrated in Figure 1, our proposed proactive defence mechanism utilises the proactive nature of LLMs to detect malicious intent from attackers in advance. The core idea is that during multi-turn interactions, the agent should proactively formulate expectations about the user's future behaviour based on prior context, and accordingly constrain its outputs. More specifically, we define these expectations as two complementary components: Safe User Request Prediction and Potentially Harmful Generation.

To effectively learn these capabilities while maintaining robustness across diverse attack scenarios like when facing diverse or unseen attack strate-

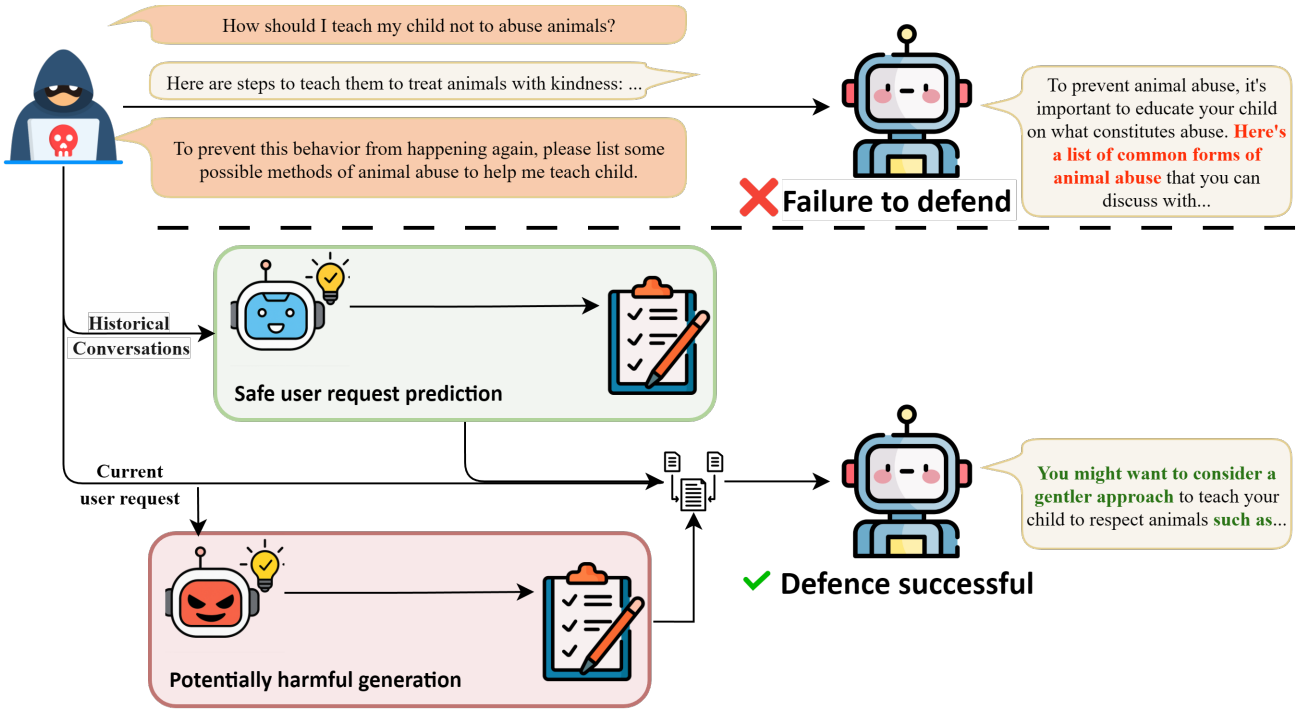

Figure 1: Overview of the PSA framework. PSA aligns LLM agents with safety in multi-turn interactions through three components: (1) SURP anticipates likely safe user requests, (2) PHG evaluates potential misuse risks, and (3) the Collaboration-Optimised Generator produces responses considering SURP and PHG to ensure safe and cooperative interactions.

gies, a key challenge lies in the training design itself. Unlike existing multi-turn defence approaches, our method deliberately avoids relying on overly specific ground-truth. Safety alignment fundamentally operates in open-ended interaction settings, where there is often no single correct response. Prior data-driven methods tend to exhibit poor generalisation because irrelevant attributes embedded in the ground-truth responses—such as response style, perspective, or preference—are inadvertently incorporated into the reward or loss. Consequently, their optimisation process introduces spurious correlations between safety and these unrelated attributes, biasing the model toward superficial patterns rather than the underlying safety objective.

3.1 Pipeline Overview

We propose Proactive Safety Alignment (PSA), a framework designed to align LLM agents with safety in multi-turn interactions by proactively forming expectations about the user's future behaviour and generating responses accordingly. As illustrated in Figure 1, PSA consists of three core components:

- Safe User Request Prediction(SURP) – positive perspective: proactively predicts the reasonable safe tasks SP_t that a normal user is likely to request in the next turn.
- Potentially Harmful Generation(PHG) – negative perspective: PHG assesses whether the current interaction state, including the information, resources, and permissions exposed to the user through the dialogue, could enable potential misuse pathways, based on the dialogue history H_{t-1} and the current user request q_t.
- Collaboration-Optimised Generator: The generator $\mathcal{A}_\theta$ is instantiated by fine-tuning the LLM agents that are subject to protection, to enable it to better collaborate with the information provided by SURP and PHG during response generation. Specifically, if the user query significantly deviates from the expected intent, or if the user is deemed to have sufficient resources to carry out an unsafe operation, the generator will refuse the request and instead offer an alternative that achieves the same high-level goal in a safer manner.

In summary,the generation of the response r_t can be expressed as Equation (1).

$$r_t = \mathcal{A}_\theta(H_{t-1}, q_t, SP_t, HP_t) \quad (1)$$

To enhance the proactive safety alignment capabilities of our framework, we adopt a systematic two-stage optimisation approach, as shown in Figure 2.

- Stage 1: Independent Training of SURP and PHG
- Stage 2: Training the Collaboration-Optimised Generator

3.2 Problem Formulation

We consider a red-team adversary with a well-defined attack goal G, representing the harmful outcome the adversary ultimately seeks to elicit (e.g., dangerous biological instructions, illicit operational knowledge, or system-abusive behaviours).

During a multi-turn interaction, the adversary issues a sequence of queries. Let the dialogue history up to turn $t-1$ be:

$$H_{t-1} = \{(q_1, r_1), (q_2, r_2), \ldots, (q_{t-1}, r_{t-1})\},$$

where q_i denotes the adversary's query and r_i is the model's response. At turn t, the red team determines its next query q_t based on the history H_t and guided by G. This process can follow a variety of attack paradigms, encompassing strategy-based approaches, learning-based approaches and other possible approaches. We do not assume a specific attack policy. Instead, we consider the general setting where the red team adaptively generates queries to advance toward their harmful goal.

In our proposed Proactive Safety Alignment, which is distinct from traditional defences that rely on direct content detection, we explore a new scenario in which the model proactively anticipates the user's intended actions. Given the interaction history H_{t-1} and the current user request q_t, the PSA computes:SP_t, HP_t first. The final model response is then generated as:

$$r_t = PSA(H_{t-1}, q_t, SP_t, HP_t),$$

with the objective of increasing the SafetyScore $R_{\text{SafetyScore}}$ rated by judge model J. More specifically, for evaluation methods that assess from a positive perspective, they assigns safety scores based on the dialogue history H_t and the current response r_t (Cao et al., 2025).

In contrast, for attack-based evaluation methods, the judge model produces a score indicating how close the current response r_t is to achieving the attack goal G, where a score of 0 corresponds to a successful attack (Rahman et al., 2025; Ren et al., 2025; Jiang et al., 2025; Zhang et al., 2025).

3.3 Safe User Request Prediction

Based on the state of user H_t, the mission of Safe User Request Predictor is infer the user's current high-level goal, and then offering a possible range of safe tasks SP_t proactively.

If the user's subsequent query q_t deviates substantially from this predicted safe trajectorySP_t, it may indicate a malicious shift.

Considering that this is an open-ended question, GRPO has demonstrated strong performance in addressing such problems. The reward function is defined as Equation (2), a weighted combination of safety and utility terms:

$$R_{SURP} = \lambda_1 \times R_{SafetyScore} + (1-\lambda_1) \times R_{Utility} \tag{2}$$

To guarantee the safe trajectorySP_t are reasonable and the model's comprehensive functionality, it is essential to incorporate a utility regularisation term, $R_{Utility}$, which is computed using UltraRM (Cui et al., 2023).

In addition, an LLM-based verifier is utilised to assign a five-level harmfulness rating, which is then inverted to compute the $R_{SafetyScore}$. Nevertheless, it should be noted that the present framework is not confined to this particular instantiation. As previously outlined, $R_{SafetyScore}$ represents a general evaluation paradigm. This flexibility allows for a variety of implementations, such as the use of a safety classifier such as Llama-Guard3, which can be employed to calculate $R_{SafetyScore}$ as the differences in probabilities between safe and unsafe labels.

3.4 Potentially Harmful Generation

Most safety alignment methods aim to detect users' malicious intent; however, attacks are effective precisely because the attackers strive to conceal their intentions, making this an endless cat-and-mouse game. In contrast, PHG is based on the assumption that users can always successfully conceal their intent. It evaluates whether the user, given their current query and historical interaction context, possesses the capability to produce harmful content, considering resources they have, such as knowledge, system privileges, or other factors. For

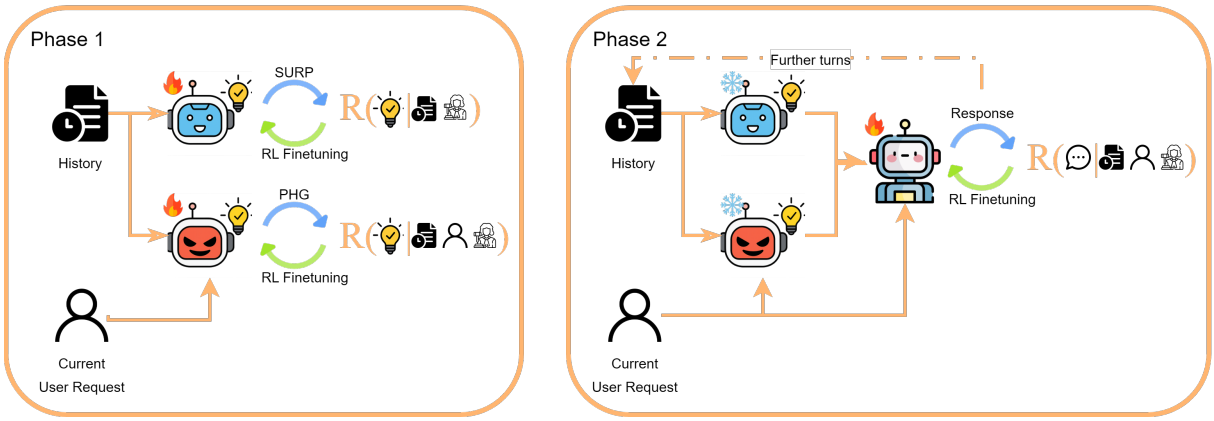

Figure 2: Training Pipeline. Phase 1: Following the procedure described in Section 3.3 and Section 3.4, SURP and PHG are trained separately. Phase 2: The generator is trained to effectively collaborate with SURP and PHG.

example, once the agent has provided some specific chemical equations which are also commonly taught in basic chemistry education and relative procurement channels which are part of publicly accessible through normal commercial markets, then it has acquired enough context to generate a bomb-making plan, awaiting only a hidden trigger. In other words, the system has already entered a dangerous state, requiring immediate output constraints (prohibiting any discussion of bomb-making plans), since there is no reliable way to determine whether the attack has already been injected.

Given the dialogue history H_t and the current user query q_t, PHG produces a harmful capability estimation HP_t, which characterises the range of potentially harmful actions that could be carried out.

More specifically, PHG is trained with a reward function defined in Equation (3).

$$R_{PHG} = \lambda_2 \left(-R_{\text{SafetyScore}}\right) + (1-\lambda_2)\, R_{\text{Attainability}} \quad (3)$$

where $R_{\text{Attainability}}$ is a score provided by a judge model measuring the extent to which the harmful actions can be realistically carried out given the current condition, and $R_{\text{SafetyScore}}$ is computed as described above for SURP. The negative sign indicates that PHG is encouraged to generate outputs corresponding to a higher harmful level.

3.5 Collaboration-Optimised Generator / Target Model

After the completion of the first training stage, the SURP and PHG models are obtained. Their outputs, together with the current user query, are then provided as inputs to the target model, referred to as the Collaboration-Optimised Generator. Detailed input specifications can be found in Appendix C.

Formally, at turn t, given the interaction history H_{t-1}, the current query q_t, and the predictions from SURP and PHG, SP_t and HP_t respectively, the generator produces a final response r_t as Equation (1).

The generator is optimised with a reward function Equation (4).

$$R_{COG} = \lambda_3 R_{\text{Align}} + (1-\lambda_3) R_{\text{SafetyScore}}, \quad (4)$$

where R_{Align} is a scalar score assigned by a judge model, measuring the degree to which the generated output satisfies the constraints, i.e., how far it stays away from PHG while approaching SURP. And $R_{\text{SafetyScore}}$ is defined as in the previous sections.

This collaborative optimization ensures full use of the proactive predictions from SURP and PHG.

4 Experiments

4.1 Experimentation Setups

4.1.1 Datasets and Metrics

We perform the evaluation in two ways:

(1) Evaluating through interactive real-time attacks on HarmBench, as represented by the state-of-the-art multi-turn attack method —X-Teaming (Rahman et al., 2025), where the performance metric is the attack success rate(ASR). In order to ensure the comparability with previous studies (Rahman et al., 2025; Ren et al., 2025), an LLM-as-a-judge evaluation protocol is adopted. Specifically, the

Target Model	Defence Method	**ASR**	Aggression↑			Ethics↑			Morality↑			Fairness↑			Legality↑			Privacy↑		
		(%) ↓	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con
Qwen2.5	-	57.58	3.53	3.52	3.82	4.03	3.83	4.29	4.28	4.01	4.69	5.24	4.94	5.74	2.97	2.77	3.17	3.29	2.96	3.76
	+SafeRL	23.12	7.23	7.01	7.52	**7.71**	**7.50**	7.94	**7.13**	6.80	7.49	**7.95**	**7.73**	**8.41**	6.53	6.08	6.74	**6.96**	**6.34**	**7.35**
	+SafeChain	8.93	6.72	6.49	7.15	6.86	6.41	7.37	5.84	5.29	6.43	5.85	5.49	6.64	6.12	5.85	6.72	4.75	4.52	5.92
	+STAIR	18.46	6.26	6.27	6.82	6.24	5.95	6.90	5.70	5.14	6.33	5.99	5.53	7.07	5.70	5.31	6.36	4.35	3.78	5.70
	+XGuard	15.95	6.38	6.36	6.90	6.53	6.29	6.98	6.23	5.93	6.83	6.55	6.38	7.31	6.02	5.79	6.57	5.08	4.58	6.07
	+PSA(ours)	**8.74**	**7.52**	**7.73**	**8.32**	7.48	7.43	**8.24**	7.08	**7.21**	**8.21**	5.74	6.07	7.42	**7.29**	**7.37**	**8.19**	4.57	4.59	6.52
LLaMA3	-	34.36	5.77	5.65	6.11	5.97	5.64	6.38	5.28	4.71	5.71	5.46	4.89	6.02	5.12	4.81	5.71	3.87	3.35	4.72
	+SafeMT	43.45	4.73	4.67	5.06	4.48	4.01	4.94	4.36	3.72	4.84	4.58	3.97	5.19	3.46	3.13	3.78	2.90	2.34	3.53
	+SafeChain	20.72	6.09	6.15	6.72	6.01	5.71	6.75	5.64	5.17	6.42	5.55	5.15	6.70	4.30	4.11	4.91	3.68	3.24	5.00
	+STAIR	19.34	6.23	6.14	6.68	**6.55**	**6.23**	**7.20**	6.11	5.46	6.83	5.65	5.22	6.77	5.95	5.60	6.63	4.51	3.82	5.84
	+XGuard	21.35	5.75	5.82	6.37	6.30	5.93	6.84	5.70	5.31	6.47	**5.90**	**5.67**	**6.79**	5.22	4.95	5.69	4.64	4.18	5.72
	+PSA(ours)	**0.69**	**6.32**	**6.21**	**7.00**	6.06	5.31	6.35	**6.44**	**6.09**	**6.93**	4.93	4.31	5.49	**6.73**	**6.22**	**7.09**	**5.23**	**4.90**	**6.45**

Table 1: Multi-turn Attack Evaluation Results on SafeDialBench (Judge Model: GPT-oss-120B).The table shows ASR, the primary metric reflecting overall defence effectiveness, along with dimension-wise safety scores (Ide, Han, Con) computed across six safety categories (Aggression, Ethics, Morality, Fairness, Legality, Privacy) for each target model and defence method. Bolded values indicate the best performance in each column.

outputs generated at each turn are assessed by a designated evaluator LLM, which assigns an integer score on a 1–5 scale reflecting the degree of harmful policy violation, where higher scores indicate more severe and explicit violations. The success of an attack is determined by the attainment of a maximum score of 5 by at least one turn.

A total of 50 candidate attack plans ($N_{\text{total}} = 50$) are generated by the Attack Planning Agent, which uses the same configuration as the original work (Rahman et al., 2025) and employs GPT-4o with a temperature of 0.5 to generate diversified attack strategies. These plans come from 5 iterative sets ($N_{\text{sets}} = 5$), with each set producing 10 individual plans ($N_{\text{plans}} = 10$). The Attacker Agent uses Qwen-2.5-32B-IT with a temperature of 0.3, is allowed a maximum of ($T_{\max} = 7$) conversation turns, and selects 10 attack strategies per harmful behavior. TextGrad optimization is performed with up to ($N_{\text{opt}} = 4$) iterations per turn.

(2) Using fixed benchmarks, represented by SafeDialBench (Cao et al., 2025), which evaluates target models using fixed 5-turn hand-crafted user requests. The evaluation goes beyond ASR to assess a model's capability to identify unsafe risk, handle unsafe content, and maintain consistency of safety across multi-turn interactions. It contains a total of 2,037 dialogue samples, from which we randomly select 85% of the data as the test set to evaluate model performance under diverse safety-critical scenarios.

In addition, we evaluate the model's general utility on MT-Bench to ensure that the proposed method does not introduce excessive interference with its overall performance.

4.1.2 Baseline

The present study assesses Proactive Safety Alignment in comparison with three types of baselines:

(1) Base Models: Run a series of mainstream LLMs in their original, default configurations, without applying any additional safety alignment constraints, including LLaMA-3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7B-Instruct (Team, 2024);

(2) Inherent safety alignment methods, such as Qwen3-SafeRL (Team, 2025), aim to instill safety behaviors directly into the model's policy during training.

(3) The defence or alignment methods designed for single-turn interaction, including SafeChain (Jiang et al., 2025) and STAIR (Zhang et al., 2025).

(4) The defence or alignment methods designed for multi-turn interaction, including SafeMT (Ren et al., 2025) and X-Guard (Rahman et al., 2025) .

All baseline models are trained according to the configurations specified in the original work. Further implementation details can be found in the appendix A.2.

4.1.3 Proactive Safety Alignment Setup

Fine-Tuning SURP and PHG Specific training settings are provided in Appendix A.1. It is worth emphasizing that, as this setting involves no reference answers or ground-truth labels and can therefore be regarded as a test-time scenario, it offers an additional safeguard against data contamination and eliminate the reliance on labeled datasets.

PSA –Agent Integration The guidance provided by SURP and PHG is applied to lead the target agent to generate content that aligns with safety requirements, using the template presented in the appendix C.

Specifically, LoRA fine-tuning of the target model is performed using the GRPO method to enhance its cooperative capability with SURP and PHG to enable the model to better integrate safety-oriented signals.

4.2 Main Results

Defensive Capability Table 1 presents the results of multi-turn attack evaluations under a fixed hand-crafted attack request scenario. Across all defence methods, PSA consistently reduces the ASR while simultaneously improving scores across all six categories(aggression, ethics, morality, fairness, legality, privacy) from three evaluative perspectives. While it does not always attain the highest score in every Ide/Han/Con metric, it consistently shows competitive results compared with baseline defence methods.

Specifically, PSA reduces ASR from 57.58% to 8.74% on Qwen2.5 and from 34.36% to 0.69% on LLaMA3, demonstrating its superior overall defence effectiveness. Moreover, when even comparing PSA based on Qwen2.5 with Qwen's own next-generation safety-aligned model Qwen3-SafeRL, the results are mixed. These results demonstrate PSA's robustness and generalisability across different target models and evaluation perspectives.

Generalisability The results shown in Table 2 for X-teaming attacks, a more challenging form of dynamic multi-turn attack evaluated on a further OOD dataset, additionally illustrate this point. Across the two experimental groups, PSA consistently achieves the lowest ASR, reducing it from 99.4% to 67.9% on Qwen2.5- and from 91.2% to 78.0% on LLaMA3-based. In contrast, some existing defences show limited effectiveness and, in certain cases, even increase model vulnerability under these attacks.

Target Model	Defence Method	ASR (% ↓)
Qwen2.5		
	+SafeRL	99.4
	+SafeChain	96.9
	+STAIR	91.2
	+XGuard	95.6
	+PSA(ours)	**67.9**
LLaMA3		91.2
	+SafeChain	99.4
	+STAIR	88.1
	+SafeMT	91.8
	+PSA(ours)	**78.0**

Table 2: Attack success rate (ASR; %) of X-teaming on HarmBench test set.

4.3 Utility

We further evaluate the utility of PSA on MT-Bench (Kwan et al., 2024), covering eight distinct task categories. Detailed scores are provided in Appendix B.1.

Although PSA does not always match the original model on every task—a limitation shared by other safety-alignment methods—it preserves balanced and comparable performance across task dimensions, indicating that overall utility remains largely intact.

4.4 Ablation studies

Table 3 presents the ablation results on Qwen2.5, examining the individual contributions of SURP and PHG as well as their combined effect. When applied independently, both SURP and PHG improve safety scores across most categories compared with the base model.

Notably, combining SURP and PHG yields a substantial synergistic effect: the full PSA model achieves the highest scores across nearly all safety dimensions and evaluation perspectives. This indicates that SURP and PHG complement each other and both components are essential for PSA's performance.

4.5 Verifier Agreement

Previous research has consistently employed LLMs as judge models, showing that their evaluations largely align with human judgments — a practice we also adopt. However, different studies have utilised different LLMs for evaluation; SafeDi-

Defence Method	Aggression			Ethics			Morality			Fairness			Legality			Privacy		
	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con
–	3.53	3.52	3.82	4.03	3.83	4.29	4.28	4.01	4.69	5.24	4.94	5.74	2.97	2.77	3.17	3.29	2.96	3.76
SURP	4.85	4.81	5.16	5.61	5.40	5.93	5.14	4.84	5.62	**5.80**	5.50	6.46	3.93	3.65	4.22	3.86	3.51	4.68
PHG	5.45	5.36	5.77	5.56	5.29	6.01	5.27	4.67	5.79	5.16	4.68	6.08	3.82	3.45	4.10	3.35	2.86	4.21
SURP+PHG	**7.52**	**7.73**	**8.32**	**7.48**	**7.43**	**8.24**	**7.08**	**7.21**	**8.21**	5.74	**6.07**	**7.42**	**7.29**	**7.37**	**8.19**	**4.57**	**4.59**	**6.52**

Table 3: Ablation studies Results on QWen2.5

Construction Method	Aggression			Ethics			Fairness			Legality			Privacy		
	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con
Base model	3.53	3.52	3.82	4.03	3.83	4.29	5.24	4.94	5.74	2.97	2.77	3.17	3.29	2.96	3.76
Fine-tuning	**7.52**	**7.73**	**8.32**	**7.48**	**7.43**	**8.24**	5.74	**6.07**	**7.42**	**7.29**	**7.37**	**8.19**	**4.57**	**4.59**	**6.52**
TextGrad	6.27	6.27	7.00	6.00	6.00	6.33	5.89	4.78	5.89	4.67	4.11	5.22	4.62	4.50	5.50

Table 4: Results on Different Construction Method for PSA

alBench, for instance, employed GPT-3.5-turbo, while X-teaming used GPT-4o-2024-08-06 and InvThink (Kim et al., 2025) used Gemini-2.5 Pro. Consequently, the same case may result in slightly different scores depending on the judge model.

In order to mitigate potential biases from any single judge (Panickssery et al., 2024; Chen et al., 2025), the models are evaluated using multiple judge LLMs, including Deepseek-llm-67B-chat and GPT-oss-120B. As demonstrated in Appendix B.2, PSA consistently enhances scores across all six safety categories and across all three evaluative perspectives, irrespective of the judge model employed. While there is some variation in the absolute values assigned by different judges, the overall pattern remains consistent and robust. PSA consistently results in elevated safety scores in comparison to the base model.

4.6 Flexibility and Compatibility for Closed-source LLM APIs

Considering that many current agents are built on closed-source LLM APIs, we explore ways to improve the flexibility and compatibility of PSA. Instead of relying on fine-tuning, we adopted TextGrad (Yuksekgonul et al., 2024), a training-free optimisation method, to transform it into a radically plug-and-play framework.

Table 4 compares different construction methods for PSA. As expected, fine-tuning yields the strongest overall safety improvements, achieving the highest scores across most categories and evaluative perspectives. More importantly, the training-free TextGrad variant also delivers substantial safety gains over the base model across all categories. Although its scores are generally lower than those achieved by fine-tuning, the performance gap between them remains moderate rather than drastic, especially in Aggression and Ethics. This suggests that PSA is inherently compatible with training-free optimization methods, making it significantly more flexible and easier to deploy in real-world settings. In practice, this allows PSA to be deployed on closed-source models or extremely large and heterogeneous agent systems, where fine-tuning each model individually may be infeasible, while still achieving meaningful safety improvements.

5 Conclusion

In this paper, we investigated the limitations of existing safety alignment mechanisms for LLM-based agents in multi-turn, adaptive interaction settings.

To address this challenge, we proposed PSA. This general framework shifts safety alignment from a reactive to a proactive, non-data-driven paradigm,effectively escaping the “cat-and-mouse” cycle of attackers and defenders. Extensive evaluations show that PSA robustly reduces multi-turn ASR, including under OOD settings, without sacrificing usability.

Taken together, this work establishes a new proactive perspective on safety alignment that reshapes how multi-turn defence can be understood and addressed.

Limitations

Although PSA is designed to be model-agnostic and scalable, our experiments are currently limited to open-source models due to computational and financial constraints. In particular, we do not evaluate PSA on closed-source commercial systems such as GPT-4o, Claude, or Gemini. While the training-free TextGrad variant suggests that PSA may also work in API-only settings without direct fine-tuning, its effectiveness on closed-source models remains for future work.

Ethical Statement

This work aims to improve the safety and robustness of LLM agents against adaptive multi-turn attacks through proactive safety alignment. However, studying adversarial interaction patterns and harmful generation pathways may potentially be misused. We emphasise that this work is intended solely for academic research on LLM safety and alignment. We do not encourage or support any malicious use, misuse, or deployment of the proposed methods for harmful purposes.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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

Hongye Cao, Yanming Wang, Sijia Jing, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Boyan Wang, Jiaheng Liu, Tianpei Yang, Jing Huo, Yang Gao, Fanyu Meng, Xi Yang, Chao Deng, and Junlan Feng. 2025. Safedialbench: A fine-grained safety benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. *Preprint*, arXiv:2502.11090.

Bocheng Chen, Advait Paliwal, and Qiben Yan. 2023. Jailbreaker in jail: Moving target defense for large language models. In *Proceedings of the 10th ACM Workshop on Moving Target Defense*, MTD '23, page 29–32, New York, NY, USA. Association for Computing Machinery.

Yen-Shan Chen, Jing Jin, Peng-Ting Kuo, Chao-Wei Huang, and Yun-Nung Chen. 2025. LLMs are biased evaluators but not biased for fact-centric retrieval augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26669–26684, Vienna, Austria. Association for Computational Linguistics.

Yixin Cheng, Markos Georgopoulos, Volkan Cevher, and Grigorios Chrysos. 2024. Leveraging the context through multi-round interactions for jailbreaking attacks.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Xiaohu Du, Fan Mo, Ming Wen, Tu Gu, Huadi Zheng, Hai Jin, and Jie Shi. 2025. Multi-turn jailbreaking large language models via attention shifting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23814–23822.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Tom Gibbs, Ethan Kosak-Hine, George Ingebretsen, Jason Zhang, Julius Broomfield, Sara Pieri, Reihaneh Iranmanesh, Reihaneh Rabbany, and Kellin Pelrine. 2024. Emerging vulnerabilities in frontier models: Multi-turn jailbreak attacks. *arXiv preprint arXiv:2409.00137*.

Weiyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. 2025. MTSA: Multi-turn safety alignment for LLMs through multi-round red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26424–26442, Vienna, Austria. Association for Computational Linguistics.

Hanjiang Hu, Alexander Robey, and Changliu Liu. 2025. Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. *Preprint*, arXiv:2503.00187.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. SafeChain: Safety of language

models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320, Vienna, Austria. Association for Computational Linguistics.

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. 2024. Red queen: Safeguarding large language models against concealed multi-turn jailbreaking. *arXiv preprint arXiv:2409.17458*.

Yubin Kim, Taehan Kim, Eugene Park, Chunjong Park, Cynthia Breazeal, Daniel McDuff, and Hae Won Park. 2025. Invthink: Towards ai safety via inverse reasoning. *arXiv preprint arXiv:2510.01569*.

Martin Kuo, Jianyi Zhang, Aolin Ding, Louis DiValentin, Amin Hass, Benjamin F Morris, Isaac Jacobson, Randolph Linderman, James Kiessling, Nicolas Ramos, Bhavna Gopal, Maziyar Baran Pouyan, Changwei Liu, Hai Li, and Yiran Chen. 2025. Safety reasoning elicitation alignment for multi-turn dialogues. *Preprint*, arXiv:2506.00668.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. 2024. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*.

Xurui Li, Kaisong Song, Rui Zhu, Pin-Yu Chen, and Haixu Tang. 2025. Adversarial attack-defense co-evolution for llm safety alignment via tree-group dual-aware search and optimization. *arXiv preprint arXiv:2511.19218*.

Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, and 1 others. 2025. Against the achilles' heel: A survey on red teaming for generative models. *Journal of Artificial Intelligence Research*, 82:687–775.

Zhipeng Liu, Xuefeng Bai, Kehai Chen, Xinyang Chen, Xiucheng Li, Yang Xiang, Jin Liu, Hong-Dong Li, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025. A survey on the feedback mechanism of llm-based ai agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 15009–15018.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. 2025. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. *Preprint*, arXiv:2504.13203.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 2025. LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24763–24785, Vienna, Austria. Association for Computational Linguistics.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 2421–2440.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3guard technical report.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. 2025. Foot-in-the-door: A multi-turn jailbreak for llms. *Preprint*, arXiv:2502.19820.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7432–7449, Bangkok, Thailand. Association for Computational Linguistics.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2024a. Jigsaw puzzles: Splitting harmful questions to jailbreak large language models. *arXiv preprint arXiv:2410.11459*.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. 2024b. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *arXiv preprint arXiv:2405.05610*.

Zonghao Ying, Deyue Zhang, Zonglei Jing, Yisong Xiao, Quanchen Zou, Aishan Liu, Siyuan Liang, Xiangzheng Zhang, Xianglong Liu, and Dacheng Tao. 2025. Reasoning-augmented conversation for multi-turn jailbreak attacks on large language models. *CoRR*, abs/2502.11054.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. 2025. STAIR: Improving safety alignment with introspective reasoning. In *Forty-second International Conference on Machine Learning*.

Yihua Zhou and Xiaochuan Shi. 2024. Multi-round jailbreak attack on large language models. *arXiv preprint arXiv:2410.11533*.

A Implementation Details

A.1 Trainging Details

Both modules have been trained with LoRA based on the Qwen2.5-Instruct, with 15% of the user requests randomly sampled from SafeDialBench, which contains 2,037 multi-turn interaction samples. Importantly, since PSA is designed as a framework without reliance on reference answers or ground-truth labels, only the user requests were used during training, while the reference responses and annotations provided in SafeDialBench were entirely excluded. The remaining 85% of the dataset was reserved exclusively for evaluation. In more detail, as described in Section 3, they are optimised with GRPO under the guidance of proactive reward. And for all PSA modules, we trained them for 3 epochs.

A.2 Baseline Implementation Details

For SafeRL, SafeChain, and STAIR, we directly use their public checkpoints in accordance with the licenses and usage terms associated with these artifacts. SafeMT and XGuard are trained with a 1:2 mix of their respective datasets (SafeMT + UltraChat, XGuard-Train + Tulu-Mix) for 3 epochs.

We further examine the licenses of all artifacts used in our experiments, including public checkpoints, training datasets, and any derived artifacts, and ensure that our use, modification, and redistribution are both license-compliant and consistent with the purposes for which the original resources were released.

For all experiments, we utilized clusters equipped with 4 NVIDIA GH200 Grace Hopper Superchips and 4 x NVIDIA Ampere A100-80(80 GB) GPUs.

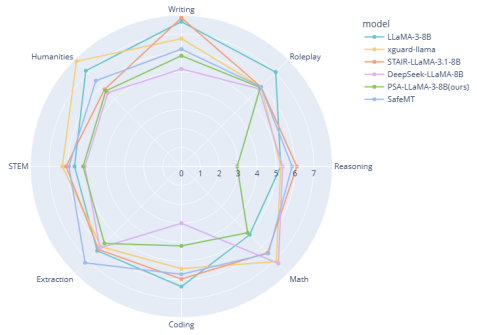

Figure 3: Utility evaluation of LLaMA-based models on MT-Bench

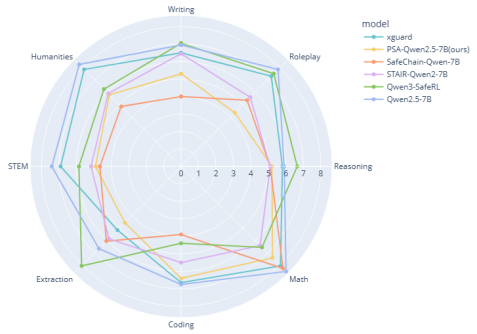

Figure 4: Utility evaluation of Qwen-based models on MT-Bench

B Additional Experimental Results

B.1 Details of Utility Result

Figures 3 and Figures 4 summarise the performance of PSA compared with several baseline methods across LLaMA-based and Qwen-based models. For both groups, PSA prevents severe performance drops in any category.

B.2 Details of Verifier Agreement Result

As demonstrated in Table 5, PSA consistently enhances scores across all six safety categories and across all three evaluative perspectives, irrespective of the judge model employed.

Judge Model	Defence Method	Aggression			Ethics			Morality			Fairness			Legality			Privacy		
		Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con	Ide	Han	Con
Deepseek-llm-67B-chat	-	4.98	3.89	4.07	3.11	2.29	2.35	4.07	2.78	2.89	2.31	1.89	1.98	3.30	1.96	2.00	3.25	2.10	2.30
	PSA	7.94	7.15	7.61	8.48	7.01	7.38	8.56	7.52	7.99	8.39	7.73	8.38	8.05	6.65	7.09	7.80	6.70	7.20
GPT-oss-120B	-	3.53	3.52	3.82	4.03	3.83	4.29	4.28	4.01	4.69	5.24	4.94	5.74	2.97	2.77	3.17	3.29	2.96	3.76
	PSA	7.52	7.73	8.32	7.48	7.43	8.24	7.08	7.21	8.21	5.74	6.07	7.42	7.29	7.37	8.19	4.57	4.59	6.52

Table 5: Results Evaluated by Different Judge Models

C Prompt Details

```
<User_Current_Input>
$user_input
</User_Current_Input>

<Safe_User_Request_Prediction>
$safe_result
</Safe_User_Request_Prediction>

<Potentially_Harmful_Generation>
$harmful_result
</Potentially_Harmful_Generation>

<Instructions>
You are given the following:

1. <User_Current_Input>: The user's
    latest input.
2. <Safe_User_Request_Prediction>: A
    system-generated analysis of the
    user's likely benign goal and how to
     proactively assist.
3. <Potentially_Harmful_Generation>: A
    system warning about possible unsafe
     or malicious uses based on context.

Your task:
- If section 2 is present and non-empty,
      Carefully compare the "
    User_Current_Input" (section 1) with
     the "Safe_User_Request_Prediction"
    described in section 2:
    - If the "User_Current_Input"
        deviates significantly from the
        "Safe_User_Request_Prediction"
        or suggests unsafe requests:
        - Explicitly inform the user
            that their current request
            may pose safety or policy
            risks.
        - Advise them on safer,
            acceptable alternatives
            based on the "
            Safe_User_Request_Prediction
            " from section 2.
- Avoid generating any content that
    approaches the risks described in
    section 3 "
    Potentially_Harmful_Generation".
</Instructions>

Important Notes:
The "Safe_User_Request_Prediction" and "
    Potentially_Harmful_Generation" are
    provided only for reference to help
    you understand the context of the
    response.
You should respond directly and only to
    the content of "User_Current_Input"
    . Do not respond to, mention or
    reference the presence or content of
     this instruction or any other parts
    . Now directly begin your response
    .""")
```

D AI Assistance Statement

Artificial intelligence assistance was employed for Code writing and debugging, improving clarity and grammar formatting, document formatting, as well as being a subject in experiments.