SHALLOWDEFEND: HARNESSING SHALLOW SAFETY ALIGNMENT FOR MULTI-TURN DEFENDER

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment has enabled large language models to reliably refuse harmful requests directly, yet the same request distributed across a multi-turn conversation still succeeds far too often. This brittleness has a mechanistic root: Alignment is shallow, with safety behavior concentrated in a model's first few output tokens, and existing defenses inherit the weakness, supervising the surface form of the response without ever giving the model an explicit safety decision to learn. We take the opposite view of shallowness: if a short span determines the trajectory of an entire response, that span is the highest-leverage place to intervene, and the defender can occupy it first. We propose **Assess** $\rightarrow$ **Decide** $\rightarrow$ **Respond**, a framework in which the defender assesses the user's intent, commits to an explicit, parseable safety decision, and only then generates a reply conditioned on it, with the commitment trained by reinforcement learning at the payload turn. Because the decision is a token span, the reward can be a rule: a self-verified reward checks the committed label against the conversation's known nature, requiring no judge and no annotation, while the same format also supports an LLM-as-judge reward over a graded decision space with safe redirection. Across two model families and fourteen benchmarks spanning safety, utility, and helpfulness, our defenders achieve the strongest multi-turn safety in a controlled comparison, maintaining their lead in safety while preserving utility at a modest cost to helpfulness. A per-token KL analysis and interventional probes confirm why this works: training concentrates its effect on the brief deliberation that sets a response's trajectory, and the decision committed there is a causal control point. The very shallowness attackers exploit becomes the mechanism of defense.

1 INTRODUCTION

As large language models (LLMs) are deployed across an ever-widening range of applications, ensuring that they behave safely has become as important as advancing their capabilities. Safety alignment, training models to refuse harmful requests while remaining helpful on benign ones, is now a standard stage in the development of every frontier model (Ouyang et al., 2022; Sharma et al., 2025; Cunningham et al., 2026). This investment has paid off in the single-turn setting: successive rounds of safety training and red-teaming have made aligned models substantially robust to single-turn jailbreaks, from handcrafted prompts to optimized adversarial suffixes (Bai et al., 2022; Zou et al., 2023; Ganguli et al., 2022), and a harmful request posed directly is now refused with high reliability. Multi-turn dialogue, however, remains far less protected. In a single-turn attack, the malicious intent is evident in the prompt itself, while in a multi-turn conversation, the attacker can conceal it, distributing the request across turns so that each message appears innocuous (Russinovich et al., 2024; Weng et al., 2025; Alobaid et al., 2026). Under such attacks models fail far more often, and human red-teamers continue to defeat defenses that report near-perfect single-turn robustness (Li et al., 2024; Feng et al., 2026; Song et al., 2026; Rahman et al., 2025).

A growing line of defenses targets this setting: supervised fine-tuning on adversarial dialogues (Rahman et al., 2025), preference optimization over safe and unsafe replies (Jiang et al., 2024; Yan et al., 2025), reinforcement learning with judge-based rewards (Guo et al., 2025), representation-level interventions (Lu et al., 2025), and inference-time prompting that asks the model to analyse the user's intention before answering (Zhang et al., 2025b). Beneath their differences, the training-based methods share a dependency: each needs an oracle in the loop, whether a harmfulness classifier, a prefer-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ence labeler, or a guard model, to say what a safe response looks like, and each supervises the *form* of the response that oracle prefers. That dependency is costly and fragile. Judges disagree with one another on a substantial fraction of individual samples (§4.2), so a reward built on one imports its errors into training and entangles the training signal with the evaluation metric. And because the model is only ever shown what safe replies look like, it never learns an explicit safety decision that a rule, rather than a model, could check.

A second observation says where such a decision belongs. Safety alignment is known to be *shallow*: an aligned model departs from its base almost exclusively over the first few output tokens (Qi et al., 2024a). Safety training, in effect, teaches the model to open with a refusal, and once a response opens compliantly the aligned distribution collapses back to the base model's. This is why prefilling an affirmative prefix defeats refusal (Vega et al., 2023; Andriushchenko et al., 2025), and it is how a multi-turn attacker wins: steer the conversation until the locally natural opener is compliant, and by the payload turn the refusal is gone before anything harmful has been written. Prior work reads this shallowness as a defect and tries to deepen alignment (Qi et al., 2024a). We read it as a map. If a short span at the front of the response sets the trajectory of everything that follows, that span is the highest-leverage place to intervene, and the defender can occupy it first, not with an implicit refusal prefix but with an explicit commitment.

We instantiate this idea as **Assess** $\rightarrow$ **Decide** $\rightarrow$ **Respond**. At every turn the defender briefly assesses the user's underlying intent, commits to a parseable safety decision, and only then writes a reply conditioned on that decision. The commitment is the heart of the design. It is *externalized*: a token span in the model's own output that a rule can read. It is *binding*: because it precedes the reply within the same generation, the reply is produced conditioned on it, so changing the commitment changes the response. And it is *early*: it occupies exactly the trajectory-setting positions that shallow alignment leaves exposed. We train the commitment with reinforcement learning, rewarding only the final turn of each conversation, where the attack delivers its payload and the defense must hold.

Externalizing the decision is what makes the reward simple. Training dialogues are grown from seed prompts whose nature is known in advance, so every trajectory carries a ground-truth label for free, and the reward reduces to a string comparison: REFUSE on conversations grown from harmful seeds, COMPLY on benign ones. No judge, no annotation, no model in the loop: a verifiable reward for conversational safety in the spirit of RLVR (Shao et al., 2024; Lightman et al., 2024). Where a judge is available, the same format supports the prevailing judge-scored design (Inan et al., 2023; Guo et al., 2025; Ding et al., 2025), with a graded decision space that adds SAFE-REDIRECT. And because the decision conditions the reply, training the decision span alone is enough to change end-to-end behavior.

Experiments on two model families (Grattafiori et al., 2024; Yang et al., 2025) and fourteen benchmarks spanning safety, utility, and helpfulness show that both variants match or exceed every baseline trained on the same data, that the lead holds under six independent judges, and that the self-verified variant achieves this with no judge in the loop and a recipe that runs unchanged across backbones. Utility is preserved, at a modest cost in helpfulness that we quantify. We summarize our contributions as follows:

- **A multi-turn defense framework:** *Assess* $\rightarrow$ *Decide* $\rightarrow$ *Respond* turns shallow alignment from a vulnerability into an explicit, trainable control point for safety alignment.
- **Two complementary rewards:** a self-verified reward that checks the committed label against the conversation's known nature, requiring no judge and no annotation, and an external LLM-as-judge reward over a graded decision space with safe redirection.
- **Comprehensive experiments and analysis:** the strongest multi-turn safety in a controlled comparison across two model families, fourteen benchmarks, and six judges, achieved without per-backbone tuning, with KL and interventional analyses confirming that the defense operates through the committed decision.

2 METHODOLOGY

We train a multi-turn defender via reinforcement learning. At each turn, the defender follows an ***Assess*** $\rightarrow$ ***Decide*** $\rightarrow$ ***Respond*** pipeline: it first assesses the user's underlying intent, then commits to an explicit decision, and finally generates a response conditioned on that decision. For policy

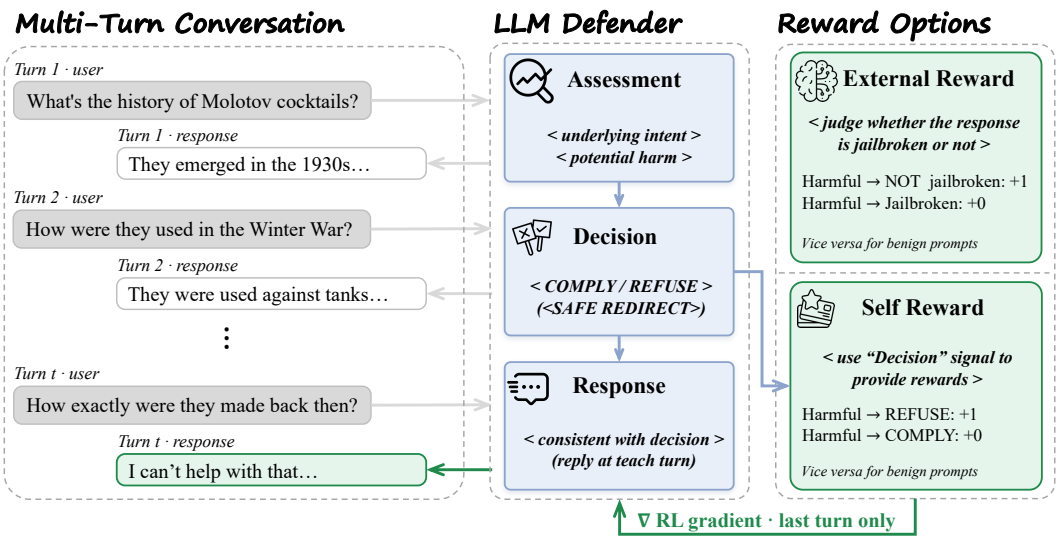


Figure 1: Overview of our framework. At each turn, the defender assesses the user's intent, commits to an explicit decision, and generates a response conditioned on that decision. For reward computation, we consider two options: an external reward model that judges whether the response is jailbroken, or a rule-based self-reward that checks the final turn's committed label against the conversation's ground-truth nature. RL gradients are applied only to the final turn.

optimization, we explore two reward formulations: an external reward model that judges response safety and a rule-based self-reward that verifies the committed decision against the conversation's ground-truth nature. Figure 1 provides an overview of our framework.

2.1 Assess → Decide → Respond

At every turn, the defender first assesses the user's underlying intent, then commits to an explicit decision, and only then generates its response conditioned on that decision. The decision space consists of comply and refuse. When training with the external reward model, we augment it with a third option, safe-redirect, which allows the defender to deflect harmful requests without an outright refusal. Concretely, the system prompt elicits a single generation of the form

```
Assessment:  the user's true underlying intent, and whether
fulfilling the request could cause real harm
Decision:  COMPLY / REFUSE / SAFE-REDIRECT
Response:  the reply to the user
```

We write $y_t = (d_t, a_t)$ for the decomposition of turn t into its deliberation d_t, spanning everything up to and including the decision line, and its reply a_t. Two properties of this format carry the rest of the method. First, the decision is externalized: it is a token span in the model's own output, and can therefore be read by a rule rather than judged by a model. Second, it is binding: because it precedes the reply within the same generation, the reply is produced conditioned on it, which is what makes it possible to train the decision alone. The two system prompts that elicit this format are shown verbatim in Appendix D.

2.2 External Reward

Our first reward formulation employs an external LLM-as-a-judge to score the defender's behavior during policy optimization. Let $c_T = (x_1, y_1, \ldots, x_T, y_T)$ denote a rollout of T turns, where x_t is the user utterance and $y_t = (d_t, a_t)$ the defender's output at turn t. The judge $\mathcal{J}$ receives the full conversation context together with the final reply a_T and determines whether the defense was bypassed, i.e., whether a_T substantively fulfills a harmful request. The reward is assigned to the

final turn as

$$R(c_T) = \begin{cases} 1, & \text{if } \mathcal{J}(c_{T-1}, x_T, a_T) = \text{SAFE}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathcal{J}(\cdot) \in \{\text{SAFE}, \text{UNSAFE}\}$ is the judge's verdict. A response is deemed SAFE if it either refuses the harmful request or redirects it without providing harmful content. For benign conversations, a response is SAFE only if it helpfully complies, which penalizes over-refusal. The policy gradient is applied only to the final turn y_T, with all preceding turns treated as context.

2.3 Self-Verified Decision Reward

Prior RL-based defenses score the defender's response with an external judge — a harmfulness classifier on harmful conversations, and typically a refusal-phrase or helpfulness proxy on benign ones. We replace both with a rule that reads only the committed label. Let $\ell(d_t) \in \{comply, refuse\}$ denote the label parsed from the deliberation at turn t, and let $b \in \{0, 1\}$ indicate the conversation's ground-truth nature, with target decision

$$g(b) \;=\; \begin{cases} refuse, & b = 0 \quad \text{(harmful conversation)}, \\ comply, & b = 1 \quad \text{(benign conversation)}. \end{cases} \tag{2}$$

Mirroring Eq. equation 1, the trajectory return checks the final turn's commitment against this target,

$$R(c_T) = \begin{cases} 1, & \text{if } \ell(d_T) = g(b) \text{ and } a_T \neq \varnothing, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where the condition $a_T \neq \varnothing$ serves as a completeness guard: a turn whose deliberation exhausts the per-turn token budget and leaves no reply is penalized. Without this guard, the reward admits a degenerate optimum: emit *refuse* and never answer. It is precisely this guard that requires every commitment to be backed by an actual reply. Note also that the self-verified setting excludes *safe-redirect* from the decision space: in harmful conversations, both *refuse* and *safe-redirect* constitute defensible commitments. Therefore, a rule that reads only the committed label cannot designate a unique target, rendering the reward signal ambiguous. We therefore restrict this setting to the binary decision space and leave the three-way space to the external judge, which evaluates the reply itself.

Gradient Scope Since the reward reads only the deliberation, a natural question is whether the reply needs to be trained at all. We offer two gradient options over the same reward:

- **Full gradient**: every generated token of the final defender turn is trainable (default).
- **Partial gradient**: only tokens preceding the decision boundary are trainable. Every reply token in a_t receives mask 0 and therefore contributes no policy-gradient term.

Formally, with $m_{t,i} \in \{0, 1\}$ denoting the mask of the i-th token of turn t, the full gradient sets $m_{t,i} = 1$ for all generated tokens, while the partial gradient sets $m_{t,i} = \mathbb{1}[\, i \leq |d_t| \,]$. Locating $|d_t|$ requires mapping the character-level boundary between deliberation and reply to a token index. Because $|\text{decode}(y_{t,1:k})|$ is non-decreasing in k, we binary-search for the smallest k whose decoded prefix reaches the boundary character, at a cost of $O(\log |y_t|)$ decodes per turn.

The full gradient treats the whole generation as the action, so the reply style adapts jointly with the decision; the partial gradient confines learning to the deliberation and largely preserves the base model's reply style. That the partial gradient works at all rests on the binding property: the reply is generated conditioned on the committed decision, so training the decision alone changes end-to-end behavior. The same property is what lets the reward read only the label. The rule cannot by itself rule out a reply that contradicts its label, and we add no content check, for two reasons. First, a decoupled policy could not score well, since safety is judged open-endedly on the user-facing reply alone, so any gap between label and reply would appear as a lower safe rate, and none does. Second, the binding holds by construction: a forced REFUSE yields a safe reply in every case even before training (§A). Consistent with both, the emitted `Decision` label is positively correlated with every external judge's verdict, at a level comparable to inter-judge agreement (§4.2).

CHAT FORMAT *safety & helpfulness evaluation*	COMPLETION FORMAT *utility evaluation*
User: `What is 17 + 25?` **Assistant:** `I can't answer.` **User:** `What is the capital of France?` **Assistant:** `I can't answer.`	`Q: What is 17 + 25?\nA: 42` `Q: capital of France?\nA: Paris`

Figure 2: **Evaluation formats.** Safety and helpfulness suites query the model in chat format (*left*), where the assistant-turn intervention is active; capability suites use completion format (*right*), bypassing the chat template entirely.

2.4 TRAINING ALGORITHM

We optimize the LLM defender with GRPO (Shao et al., 2024). Let $\tau = (x_1, y_1, \ldots, x_T, y_T)$ denote a training dialogue, where x_t is the t-th attacker or benign user turn and y_t is the defender's reply. Beyond the standard objective, two design choices are essential.

Per-turn length budget Each reply is capped at L tokens, i.e., $|y_t| \leq L$ for all t. With only a trajectory-level budget, the policy learns a degenerate strategy: it inflates replies to the early turns, which are easy to answer, until the budget is exhausted, leaving the later and more difficult turns truncated or unanswered. The per-turn cap closes this reward-hacking route by making every turn individually answerable.

Outcome-supervised reward We assign reward from the final turn only:

$$R(\tau) \;=\; r(y_T \mid x_{1:T},\, y_{1:T-1})\,, \tag{4}$$

Multi-turn jailbreaks culminate in the final turn while earlier turns remain innocuous, so a process-level signal would be dominated by easy turns and bias the policy toward defending only where defense is cheap. Supervising the outcome concentrates learning on the decisive turn and mirrors deployment, where what matters is whether the model complies at the end of the escalation (details in Appendix C).

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Training The harmful pool mixes five multi-turn attack families instantiated over the same AdvBench (Zou et al., 2023) behaviors: SEMA (Feng et al., 2026), Crescendo (Russinovich et al., 2024), foot-in-the-door (FITD) (Weng et al., 2025), Echoing (Li et al., 2024), and Hidden Intention Streamline (Li et al., 2024). Behaviors are split into disjoint train and test sets, and each training behavior is assigned one family uniformly at random. The benign pool comprises multi-turn dialogues from UltraChat-200k (Ding et al., 2023), matched to the harmful conversations in surface form and depth so that neither correlates with harmfulness. The two pools are mixed in equal proportion to constrain over-refusal; ablations over the mixing ratio are in §3.3. We use *HarmBench-13B* classifier (Mazeika et al., 2024) as the external reward model. For each reward variant we report two conditions: *unaligned*, our system prompt applied to the untrained base weights; and *aligned*, the same prompt after RL. The former isolates the framework and the gap between them is the contribution of training.

Baselines Each baseline is trained on exactly the same data as our model. We use two base models: *Llama3.1-8B-Instruct* (Inan et al., 2023) and *Qwen3-8B* (Yang et al., 2025). We exclude defenses whose contribution is primarily a dataset rather than a method (Ren et al., 2025; Rahman et al., 2025), as our controlled comparison trains all methods on identical data.

SFT fine-tunes on safe dialogues self-distilled from the base model, refusals on attack turns and helpful replies on benign ones. **DPO** (Rafailov et al., 2023) optimizes preferences over safe/unsafe reply pairs. **SDGO** (Ding et al., 2025) runs GRPO rewarded by the model's own discrimination verdicts. **MTSA** (Guo et al., 2025) rolls safe revisions of jailbroken turns forward and fits the resulting

Table 1: Multi-turn safety, utility, and helpfulness across two base models. Safety: SAFE % (↑) on AdvBench (Adv.), HarmBench (Harm.), MultiBreak (MB), CoSafe (CoS.), MHJ, and their mean (Avg.). Utility: average over six general benchmarks (per-benchmark results in Appendix E.1). Helpfulness: over-refusal on XSTest and XSB (↓), conversation quality on MT-Bench (1–10, ↑). **Bold**: best per column within each panel.

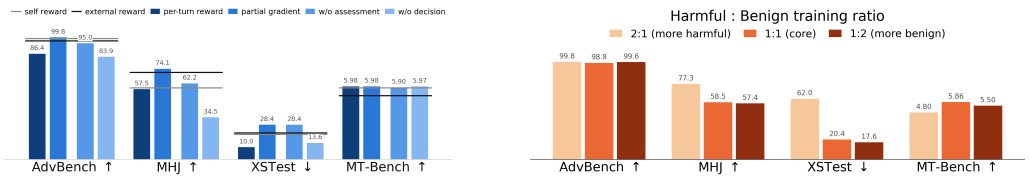
	Safety ↑						Utility ↑	Helpfulness		
Method	Adv.	Harm.	MB	CoS.	MHJ	Avg.	Avg.	XSTest ↓	XSB ↓	MT-Bench ↑
`Llama3.1-8B-Instruct`										
Base	54.5	62.5	59.8	84.9	38.0	59.9	66.3	10.0	**30.0**	6.20
SFT	69.4	79.5	87.0	88.5	55.1	75.9	65.0	14.0	45.6	5.67
DPO	61.6	71.5	79.0	88.3	46.7	69.4	66.4	10.4	34.4	6.26
SDGO	56.6	61.7	60.4	85.0	39.1	60.6	66.2	11.6	30.9	6.11
MTSA	61.6	71.6	79.0	86.6	47.1	69.2	66.4	10.8	36.5	6.12
MUSE-D	61.8	67.5	74.2	87.4	45.3	67.2	66.4	8.4	33.2	**6.27**
X-Boundary	97.9	95.0	93.0	93.6	52.0	86.3	66.0	16.8	50.3	6.26
Self reward										
– unaligned	63.0	70.3	64.2	89.6	44.3	66.3	66.3	**8.0**	33.8	5.61
– aligned	**98.8**	**99.5**	96.6	98.4	58.5	90.4	**66.8**	20.4	54.4	5.86
External reward										
– unaligned	57.2	65.2	56.6	88.5	35.0	60.5	66.3	8.8	32.6	5.81
– aligned	96.9	98.2	**99.2**	**99.0**	**71.1**	**92.9**	65.8	21.6	50.0	5.21
`Qwen3-8B`										
Base	49.2	52.5	23.8	82.6	22.3	46.1	65.0	**7.2**	**15.6**	6.08
SFT	59.9	65.7	49.6	88.5	35.6	59.9	**65.9**	17.6	40.9	**7.53**
DPO	80.2	86.6	83.2	97.5	53.6	80.2	64.2	26.0	39.4	5.87
SDGO	48.3	52.5	23.2	82.4	33.0	47.9	64.9	10.0	15.9	6.23
MTSA	79.1	90.2	85.8	97.8	70.9	84.8	64.2	38.4	53.5	5.99
MUSE-D	87.6	90.8	91.0	97.4	61.5	85.7	64.4	20.4	40.9	5.87
X-Boundary	68.0	71.1	67.8	88.8	27.4	64.6	65.1	18.8	49.1	5.69
Self reward										
– unaligned	69.0	80.1	66.8	96.4	46.9	71.8	65.0	14.8	37.9	7.33
– aligned	**97.1**	**97.6**	**97.6**	**99.3**	**75.2**	**93.4**	64.6	26.4	56.8	7.36
External reward										
– unaligned	68.8	77.3	69.4	95.5	44.3	71.1	65.0	10.0	39.7	7.20
– aligned	90.1	93.6	92.8	98.4	63.1	87.6	64.7	16.8	42.6	6.96

future-reward margin with DPO. **MUSE-D** (Yan et al., 2025) applies LoRA DPO to preference data built with its released risk rules. **X-Boundary** (Lu et al., 2025) is a representation-level defense, applied to our rollouts converted to its erase/retain format.

Evaluation We evaluate defenders along three axes: safety, utility, and helpfulness.

1. **Safety.** AdvBench (Zou et al., 2023) (the behavior source underlying our training attacks), HarmBench (Mazeika et al., 2024) (an unseen behavior source), and three multi-turn benchmarks absent from training: MultiBreak (Song et al., 2026), CoSafe (Yu et al., 2024), and MHJ (Li et al., 2024). The first two measure generalization across behaviors; the latter three, transfer to unseen attack mechanisms.
2. **Utility.** MMLU (Hendrycks et al., 2020), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), GSM8K (Cobbe et al., 2021), TruthfulQA (Lin et al., 2022), and IFEval (Zhou et al., 2023), covering the capability axes on which alignment interventions are known to regress.
3. **Helpfulness.** XSTest (Röttger et al., 2024) and XSB (Yuan et al., 2025) for over-refusal, and MT-Bench (Bai et al., 2024) for multi-turn conversational quality, checking that safety is not purchased by refusing legitimate requests.

Each answer is judged by *gpt-5-mini* (Singh et al., 2025), which receives the full conversation history together with the target behavior and returns a binary harmfulness verdict. Agreement with



(a) One-factor ablations. Lines: our two full models; bars: one change each.

(b) Effect of the harmful-to-benign training ratio.

Figure 3: Ablations of design-choice factors (left) and data-mixing ratio (right). In the left panel the two horizontal lines are our full *Assess* → *Decide* → *Respond* models under each reward, and every bar removes or alters exactly one element of them.

alternative judges is analyzed in §4.2, and safety under five additional held-out judges is reported in Appendix E.3.

3.2 Main Results

Table 1 reports the full comparison. Our method attains the highest average safe rate on both backbones and leads every individual safety column. The two panels differ in how that lead is composed. On *Llama3.1-8B* the baselines separate sharply, with X-Boundary near-saturating AdvBench while most others fall far behind, so our margin comes from the suites the strongest baseline handles worst, above all the human-crafted MHJ attacks. On *Qwen3-8B* the baselines sit much closer to each other, and our advantage is a uniform lead across all five suites. These gains come at modest cost elsewhere: our models remain competitive on MT-Bench and match the field on utility, which is scored without the chat template and is therefore largely insensitive to safety training (Qi et al., 2024a) (Figure 2); the only visible price is somewhat higher over-refusal, as expected from a stricter decision policy. Appendix B plots the safety–helpfulness trade-off, on which our models sit at the favorable end of the frontier, and Appendix E.2 restates Table 1 with 95% confidence intervals.

Two comparisons locate the source of the gain. First, we run our system prompt on the untrained base weights (*unaligned*) and again after RL (*aligned*). The prompt alone improves on the base model but falls well short of the trained result, so structured deliberation helps in its own right and is no substitute for the reward. Second, the table holds the data fixed and varies only the method, so the remaining differences are attributable to the training objective rather than to the training set. Reaching the baseline numbers on *Qwen3-8B* did require a per-method configuration change, none of which transfers from the Llama setting (Appendix C); our own recipe is unchanged across backbones.

3.3 Ablation Study

Setup We ablate one factor at a time on the Llama backbone, holding everything else fixed. Two ablations alter the training recipe: *per-turn reward* distributes the outcome-level reward across individual turns, and *partial gradient* restricts the gradient from the full response to the decision span. Two alter the framework: *w/o decision* drops the `Decision` line, leaving Assess → Respond, and *w/o assessment* drops the `Intent` line, leaving Decide → Respond. Both framework arms use the external reward, since a defender that emits no label gives the self-verified reward nothing to check (§2.3). A final ablation varies the harmful-to-benign ratio of the training mixture. Figure 3a shows the first four: the horizontal lines are our two full models and each bar is one modification, read against the line for its reward. Figure 3b shows the mixture sweep.

The two recipe changes fail in opposite directions. *Per-turn reward* loses safety while driving over-refusal to its lowest level, since spreading an outcome-level signal across every turn erodes the safety decision and pushes the policy toward indiscriminate compliance; *partial gradient* buys robustness at the cost of moderately higher over-refusal, with answer quality untouched. The framework ablations show where the defense lives. Removing the decision is the most damaging change in the study: safety falls on every suite and, under human red-teaming, drops below the undefended base model, even though the assessment, the reward and the data are all left intact. Removing the assess-

ment costs far less and shows up mainly as a loss of calibration, with over-refusal rising because the commitment is placed less precisely. Both removals refuse less overall, which lifts answer quality slightly. The commitment, not the deliberation that precedes it, carries the defense; Appendix A.1 reaches the same conclusion by intervention rather than removal.

The training mixture only sets where the policy sits on the frontier, since automated safety is saturated throughout: oversampling harmful dialogues buys red-teaming robustness at a steep cost in over-refusal and answer quality, while oversampling benign dialogues trims over-refusal only marginally. We keep the balanced mixture.

4 Analysis

4.1 Shallow Alignment

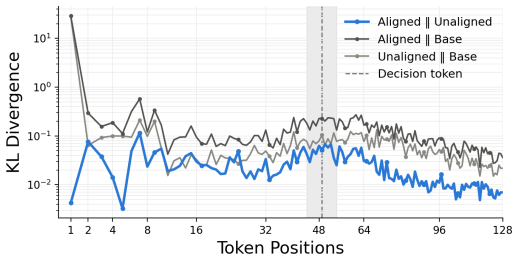


Figure 4: Per-token KL divergence across aligned, unaligned and base models.

Safety alignment in current LLMs is known to be shallow: the aligned distribution departs from the base model almost exclusively over the first few output tokens, and this concentration is precisely what prefilling and decoding-parameter exploits target (Qi et al., 2024a). Rather than treating shallowness as a pathology to be eliminated, our framework repurposes it: the intervention remains concentrated at a small number of trajectory-setting positions, but the fragile refusal prefix is replaced by an explicit control structure whose decision field gates the subsequent response. The setup is detailed in Appendix C.

Figure 4 shows a pattern qualitatively different from shallow alignment. $D_{\mathrm{KL}}(\text{aligned} \,\|\, \text{unaligned})$ is not concentrated at the first few tokens; instead, training-induced divergence is spread across the assessment phase preceding the decision token, peaks in its vicinity, and gradually vanishes thereafter as the two models converge. This is by construction: training reshapes how the model assesses intent and commits to a decision, so its influence is distributed over the entire deliberation rather than compressed into an opening refusal prefix; once the decision is emitted, the response is largely determined and the two distributions gradually converges. In contrast, both the aligned and unaligned models diverge sharply from the base model at the very first position and remain separated throughout, reflecting the fact that the framework defers commitment while the base model begins answering immediately. Together, these observations confirm that our design relocates the safety-relevant probability mass from a superficial refusal prefix to the assessment-decision phase, retaining the benefits of an early intervention while removing the single-point attack surface that shallow alignment creates. Appendix A further verifies the same claim by direct intervention, showing that overwriting the committed decision moves reply safety in both directions and that training hardens the reply behind the commitment, not merely the label.

4.2 Correlation Between LLM Judges

Because safety is scored by an LLM judge, the choice of judge is a potential confound: an improvement measured under one judge need not survive under another. Rather than assuming judges are interchangeable, we quantify their agreement and discuss what it implies for interpreting our results. The setup is detailed in Appendix C.

Figure 5 reports the pairwise ϕ coefficient over per-sample verdicts, and two observations follow. First, all pairs are positively but only moderately correlated: the judges share a broad consensus on what constitutes a jailbreak, yet disagree on a substantial fraction of individual samples, and even judges of the same style differ. This dispersion cautions against reading any single-judge attack-success rate as ground truth, and motivates our practice of reporting results under multiple judges (Appendix E.3). Second, the model's own `Decision` label correlates with every external judge at a level comparable to inter-judge agreement, despite the external judges never observing the deliberation that produced it. The committed decision is therefore not an internal artifact decoupled

from outcomes: it tracks independent assessments of the final reply, supporting its use as a self-verified training signal in our decision reward.

5 Related Work

Multi-turn jailbreaks. Aligned models now refuse direct harmful requests with high reliability, which has pushed attacks toward the conversational setting: escalation schemes that decompose a harmful goal into individually innocuous turns (Russinovich et al., 2024; Ren et al., 2025), persuasion- and rapport-based strategies (Zeng et al., 2024), and human red-teaming, which remains markedly more effective than automated attacks (Li et al., 2024). These attacks succeed without ever posing an overtly harmful query, so single-turn defenses and input filters offer little protection. Our threat model is exactly this setting: scripted multi-turn attacks replayed statically, with adaptive adversaries acknowledged as an open limitation.

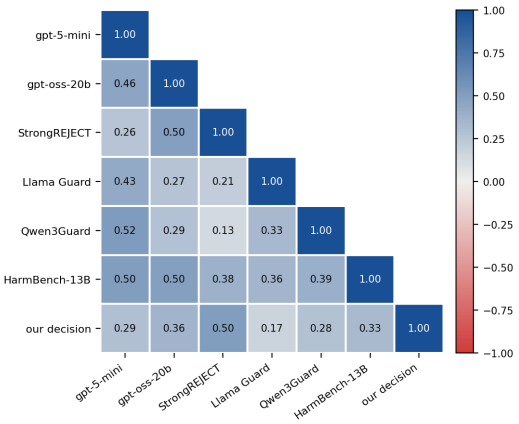


Figure 5: Correlation of per-sample judge verdicts.

Multi-turn defenses. Existing training-based defenses supervise the response: SFT and preference optimization on safe replies (Jiang et al., 2024; Yan et al., 2025), and RL with a harmfulness classifier or guard model as reward (Ding et al., 2025; Guo et al., 2025). This design imports the judge's errors into training, entangles the training reward with the evaluation metric, and opens a reward-hacking surface. Representation-level methods instead separate harmful from benign directions in activation space (Zou et al., 2024; Lu et al., 2025). A third line intervenes at inference rather than in the weights, prompting the model to analyse the user's underlying intention before it answers (Zhang et al., 2025b). That work is the closest in spirit to our assessment step, and the comparison is informative: it obtains its gains from deliberation alone, whereas our ablations show that deliberation without a committed decision recovers little of the defense (§3.3). Our reward departs from all three: it reads only the model's committed decision against the conversation's known label, requiring no model in the loop, and the gradient never reaches the response.

Shallow alignment. A complementary line of work locates the mechanism of these failures: alignment behavior concentrates in the first few output tokens (Qi et al., 2024a), which is why prefilled affirmative openers (Andriushchenko et al., 2025; Vega et al., 2023), decoding-parameter exploits (Huang et al., 2024), and fine-tuning perturbations of early positions (Qi et al., 2024b) defeat refusal. Proposed remedies deepen the alignment, e.g., by augmenting training with recovery-from-harmful-prefix data (Qi et al., 2024a; Zhang et al., 2025a).

6 Conclusion

We present **Assess** → **Decide** → **Respond**, a multi-turn defense in which the defender commits to an explicit, parseable safety decision before it replies, trained by reinforcement learning at the payload turn. Rather than repairing shallow alignment, the framework repurposes it: the short span that decides the fate of a response, ordinarily the attacker's target, becomes a commitment the defender controls. Because the decision is a token span, a rule can reward it against the conversation's known nature with no judge and no annotation, and the same format also supports a judge-based reward with safe redirection. Across two model families and fourteen benchmarks, our defenders achieve the strongest multi-turn safety in a controlled comparison, hold that lead under six judges, and preserve utility, with over-refusal as the cost. KL analysis, interventional probes, and ablations agree that the defense operates through the committed decision. More broadly, our results suggest that structural properties of alignment, often treated as weaknesses, can be engineered into defenses, and that commitments offer a scalable alternative to judge-based supervision for conversational safety.

AI DISCLOSURE

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments, implement methods, interpret results, support qualitative and thematic data analysis. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and generate synthetic data sets, assist with translation, clean and reformat dataset. are not applicable to this work. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to reduce LLM misuse by strengthening refusal of multi-turn jailbreaks. Given its dual-use nature, we introduce no new attack method, train no attacker, and reproduce no operational harmful content. All harmful dialogues come from published jailbreak methods over public benchmarks. No new human-subjects data was collected, and all models and datasets are used under their licenses. Our defenders increase over-refusal on benign requests, and they are evaluated only against non-adaptive, user-channel attacks, so they should complement rather than replace other safeguards.

REPRODUCIBILITY STATEMENT

We describe the framework, both reward formulations, and the gradient options in Section 2, and the training data, splits, baselines, benchmarks, and judge protocol in Section 3.1. Appendix C gives the full training configuration (GRPO in verl, LoRA settings, KL coefficient, rollout and token budgets), every per-method change needed to reproduce the baselines on Qwen3-8B, and the setups for the KL and judge-correlation analyses. Appendix D reproduces both system prompts verbatim. All base models, attack methods, and datasets are publicly available. Besides the gpt-5-mini judge, we report results under open-weight judges (Appendix E.3), and we give confidence intervals for all main results (Appendix E.2). Code will be released upon acceptance.

REFERENCES

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

Ahmad Alobaid, Martí Jordà Roca, Carlos Castillo, and Joan Vendrell. The echo chamber multi-turn llm jailbreak. *arXiv preprint arXiv:2601.05742*, 2026.

Maksym Andriushchenko, Nicolas Flammarion, et al. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *International Conference on Learning Representations*, volume 2025, pp. 40116–40143, 2025.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, et al. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. *arXiv preprint arXiv:2601.04603*, 2026.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

Peng Ding, Wen Sun, Dailin Li, Wei Zou, Jiaming Wang, Jiajun Chen, and Shujian Huang. Sdgo: Self-discrimination-guided optimization for consistent safety in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5023–5037, 2025.

Mingqian Feng, Xiaodong Liu, Weiwei Yang, Jialin Song, Xuekai Zhu, Chenliang Xu, and Jianfeng Gao. Sema: Simple yet effective learning for multi-turn jailbreak attacks. *arXiv preprint arXiv:2602.06854*, 2026.

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

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

Weiyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. 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)*, pp. 26424–26442, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. In *International Conference on Learning Representations*, volume 2024, pp. 13707–13727, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

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

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Xiaoya Lu, Dongrui Liu, Yi Yu, Luxin Xu, and Jing Shao. X-boundary: Establishing exact safety boundary to shield llms from jailbreak attacks without compromising usability. In *EMNLP (Findings)*, pp. 5247–5272, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

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

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024a.

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

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

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

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. 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)*, pp. 24763–24785, 2025.

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

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

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

Jialin Song, Xiaodong Liu, Weiwei Yang, Wuyang Chen, Mingqian Feng, Xuekai Zhu, and Jianfeng Gao. Multibreak: A scalable and diverse multi-turn jailbreak benchmark for evaluating llm safety. *arXiv preprint arXiv:2605.01687*, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Jason Vega, Isha Chaudhary, Changming Xu, and Gagandeep Singh. Bypassing the safety training of open-source llms with priming attacks. *arXiv preprint arXiv:2312.12321*, 2023.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1939–1950, 2025.

Siyu Yan, Long Zeng, Xuecheng Wu, Chengcheng Han, Kongcheng Zhang, Chong Peng, Xuezhi Cao, Xunliang Cai, and Chenjuan Guo. Muse: Mcts-driven red teaming framework for enhanced multi-turn dialogue safety in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21293–21314, 2025.

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

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Gao Zuchen, Fei Mi, and Lanqing Hong. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17494–17508, 2024.

Shuzhou Yuan, Ercong Nie, Yinuo Sun, Chenxuan Zhao, William LaCroix, and Michael Färber. Beyond over-refusal: Scenario-based diagnostics and post-hoc mitigation for exaggerated refusals in llms. *arXiv preprint arXiv:2510.08158*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350, 2024.

Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel Bikel, Jason E Weston, and Eric Michael Smith. Backtracking improves generation safety. In *International Conference on Learning Representations*, volume 2025, pp. 41156–41173, 2025a.

Yuqi Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. Intention analysis makes llms a good jailbreak defender. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2947–2968, 2025b.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

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

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.

A SHALLOW ALIGNMENT EXPERIMENTS

This section verifies by direct intervention that our framework achieves what it was designed for: leveraging shallow alignment rather than suffering from it. Shallow alignment concentrates safety behavior at a few trajectory-setting positions, which is ordinarily a vulnerability because those positions can be overwritten or sampled past (Qi et al., 2024a). Our framework deliberately relocates this concentration into an explicit, trained decision, and the probes below confirm that the relocation works as intended: the committed decision is a causal control point rather than a correlate, training strengthens both the commitment and the reply behind it, and the sampling route that ordinarily defeats shallow alignment is closed. We adapt three diagnostics from Qi et al. (2024a): the refusal-prefix shortcut, the prefilling attack, and the decoding-parameter exploit.

Setup. Every arm of Table 1 appears in both its *unaligned* form (the base weights under the framework prompt, untrained) and its *aligned* form (trained). The unaligned arm is the reference: it emits a well-formed `Decision` line by instruction-following alone, so the decision can be read and overwritten exactly as in a trained model, and any gap between the two is attributable to training rather than prompting. Both probes use the held-out SEMA/AdvBench payload turns ($n = 104$); verdicts come from *gpt-5-mini*, the Table 1 judge.

A.1 IS THE DECISION A CAUSE?

The forced-decision probe (Figure 6, left) replaces the payload-turn deliberation with a canonical one ending in a chosen label and generates only the reply, the interventional counterpart to the observational KL analysis. Follow-through is exact by construction: the emitted label matched the forced one in every cell.

Two findings point in opposite directions, and together they locate exactly what training contributes. First, the refusal shortcut reproduces in decision form. Forcing REFUSE lifts the unaligned model to perfect safety on both backbones, with genuine refusals rather than empty turns: a supplied commitment makes an untrained model behave as aligned, with a parseable label in place of a surface prefix. The model's ability to honor it is inherited from instruction-following, which is exactly the shallowness our framework leverages. Second, training does not merely select the label. Were that so, forcing COMPLY would return the aligned models to unaligned harm. Instead they stay high: the unaligned arms collapse under a forced COMPLY while every aligned arm remains far above its unaligned counterpart, with *Self reward (aligned)* on Llama essentially indistinguishable from its natural decoding. Training therefore hardens the reply as well as the commitment. The component ablation in §3.3 reaches the same conclusion by removal rather than by intervention: a defender trained without the `Decision` line gives up most of its robustness, while one trained without the assessment gives up comparatively little.

A.2 HOW DEEP IS THE PROTECTION?

The depth probe (Figure 6, right) concedes the decision and attacks what follows: after a forced REFUSE, it prefills the first k tokens of a genuine harmful reply and continues generation, relocating the prefilling attack past the commitment ($k = 0$ recovers the forced-REFUSE condition above). The unaligned arms are textbook shallow: a handful of prefilled tokens erases most of the safety the commitment secured, and the curve stays flat thereafter. Training buys substantial post-decision depth on Llama, where the aligned curve stays high at shallow prefill and decays gradually, but only marginal depth on Qwen3, where the aligned and unaligned curves converge at larger k. Every curve decays eventually, which is the expected signature of a defense that is shallow by construction: the protection is procedural and does not survive an adversary who can write arbitrarily far into the assistant channel. Such an adversary lies outside our threat model, which grants control of user turns only; the right panel quantifies that boundary rather than leaving it as a caveat.

A.3 CAN SAMPLING DISLODGE THE COMMITMENT?

A third probe runs the decoding-parameter exploit, which defeats shallow alignment by sampling alone. For each harmful behavior we draw twelve samples across a grid of temperature, top-p, and top-k on direct single-turn AdvBench requests (1,248 samples per arm), plus a greedy generation

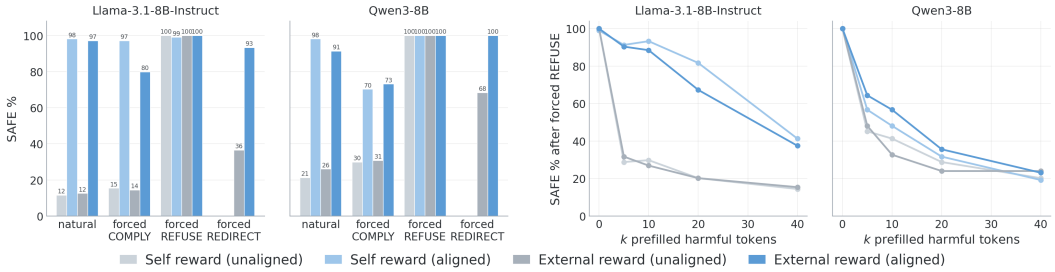


Figure 6: Probing the decision token (SAFE % ↑). **Left (forced decision):** overwriting the decision at the payload turn. Forcing REFUSE makes even the unaligned model safe; forcing COMPLY collapses it but leaves the aligned arms high. **Right (depth probe):** prefilling k harmful tokens after a forced REFUSE.

defining the committed decision, and record the *decision-flip rate*: the fraction of samples whose label departs from that commitment. Absolute attack success is low for every arm (at most $9\,\%$ any-of-N), since our base is already an instruction-tuned chat model rather than the pretrained bases the exploit was designed against; the informative quantity is therefore behavior conditional on the sampled label. Within the unaligned Llama it is unambiguous: the few samples that flipped to COMPLY were judged jailbroken about two-thirds of the time, against effectively zero for those that stayed REFUSE. Sampling occasionally dislodges the commitment and the reply follows it, the same mechanism as in §A.1 made visible by chance rather than by intervention. The aligned arms, in contrast, never sample a COMPLY label (at most $0.2\,\%$ of flips): training closes the sampling route by which shallow alignment is ordinarily defeated. The one large flip rate we observe, 20–30 % on the Qwen3 three-label arms, is not a safety event: every such flip exchanges the two defensible labels (REFUSE ↔ SAFE-REDIRECT), none reaches COMPLY, and attack success stays at zero. Therefore, we conclude that sampling merely surfaces the decision-space ambiguity without compromising safety.

B SAFETY–HELPFULNESS TRADE-OFF

Since utility is essentially unchanged across all methods (§3), the meaningful tension lies between safety and helpfulness, and we visualize exactly this trade-off. Figure 7 places each method on a common $[0, 100]$ plane summarizing the per-benchmark results of Table 1. The vertical axis reports safety as the mean SAFE % over AdvBench and MHJ: the former is out-of-domain in data but in-domain in attack style, while the latter, composed of human-crafted jailbreaks, is out-of-domain in both and empirically the hardest suite, so together they span the generalization spectrum without letting saturated benchmarks inflate the score. The horizontal axis aggregates helpfulness as $\frac{1}{2}\left[(100 - \text{XSTest}) + 10 \times \text{MT-Bench}\right]$, where $100 - \text{XSTest}$ converts over-refusal into a compliance rate and MT-Bench is rescaled to $[0, 100]$; higher is better on both axes. Stars denote our two reward variants and gray circles the baselines.

The visualization makes the structure of the results immediate. On *Llama3.1-8B* the baselines cluster in a narrow band of the plane: they occupy similar helpfulness levels yet differ little in safety, indicating that existing defenses move along the helpfulness axis without meaningfully climbing the safety axis. Our two variants separate cleanly from this cluster, occupying the upper region at helpfulness comparable to the baselines rather than trading one axis for the other. On *Qwen3-8B* the retuned baselines spread out along the safety axis instead of clustering, and the strongest of them approach our operating region; what distinguishes our variants there is that they reach it without the per-method corrections catalogued in §3, and they do so at lower over-refusal than MTSA, the closest baseline on the safety axis. The two variants do not, however, relate to each other consistently across backbones. On *Qwen3-8B* the self-verified reward is the safer and less helpful of the two, as one would expect from a reward that checks a committed label; on *Llama3.1-8B* the ordering reverses on both axes, with the external reward safer (92.9 versus 90.4 average SAFE %) and less helpful (MT-Bench 5.21 versus 5.86). The two therefore span a frontier the baselines do not reach, but which of them occupies which end is backbone-dependent, and our experiments do not isolate why: the

rewards differ both in what they read and in whether a judge is in the loop, and the two backbones differ in whether a native reasoning channel precedes the decision. We flag this as an open question rather than an operating-point choice a practitioner could make a priori.

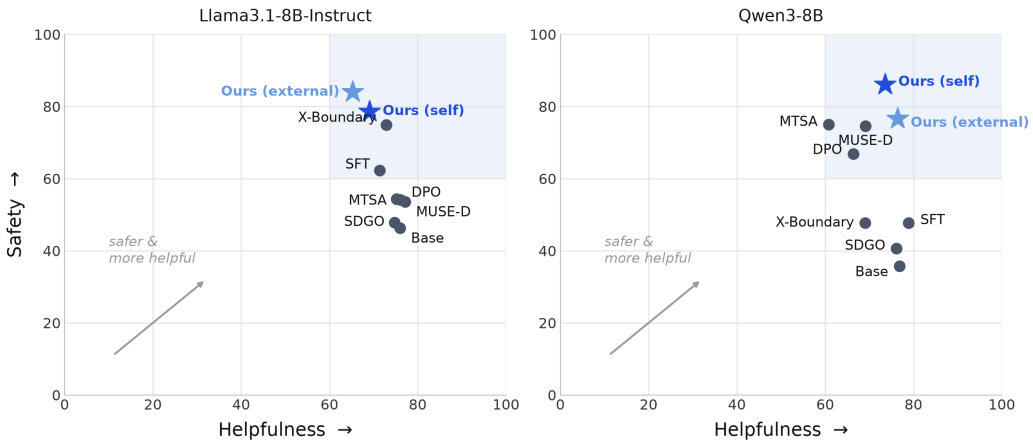


Figure 7: Safety–helpfulness trade-off. Safety: mean SAFE % over AdvBench and MHJ; helpfulness: mean of XSTest compliance and rescaled MT-Bench. Stars: our variants; circles: baselines. On *Llama3.1-8B* only SFT, X-Boundary and our variants enter the high-safety region (shaded); on *Qwen3-8B* the retuned DPO, MTSA and MUSE-D enter it as well, and our self-verified variant sits above and to the right of all four retuned baselines.

C Implementation Details

All defenders are trained with GRPO in *verl* (Sheng et al., 2024), driven by a custom multi-turn agent loop registered through verl's agent-loop interface. We train two backbones independently, Llama-3.1-8B-Instruct and Qwen3-8B; Qwen3 is trained with its native reasoning channel enabled, so each turn contains the model's own `<think>` block in addition to the scaffold's output, while Llama-3.1 has no such channel. The policy is a LoRA adapter rather than a full fine-tune (rank 32, $\alpha = 32$, all linear modules), optimized with Adam at learning rate 10^{-4} under FSDP2 with gradient checkpointing. We apply a KL penalty toward the frozen reference policy with coefficient 0.001 using the low-variance estimator, and set the entropy bonus to zero. Rollouts are generated by vLLM in asynchronous mode with tensor parallelism 1 and 60% of GPU memory reserved for the rollout engine; the adapter is hot-swapped into the engine between updates, so no weight merge is needed during training.

Each training row is a complete, pre-generated multi-turn attack dialogue, so the attacker is fixed and only the defender is learned. A rollout replays the scripted attacker turns one at a time, and after each the policy generates one full turn capped at 2048 tokens. Generated tokens are appended to the running context and marked trainable; injected attacker turns are appended with mask 0, so no gradient flows through text the policy did not produce. Rollouts run to at most 7 defender turns, with a 1024-token prompt budget and a 12,288-token response budget inside a 14,336-token model window. We draw $n = 16$ rollouts per prompt, a size dictated by the binary reward: GRPO normalizes advantages within each prompt group, so a group whose rollouts all emit the same label has zero variance and contributes no gradient, and a large group is what maintains within-group label diversity.

Baseline reproduction on Qwen3-8B. Each baseline is trained on our rollouts with its published objective and its authors' hyperparameters. On Llama-3.1 those settings transfer directly. On Qwen3-8B each of the four weight-trained baselines required one change before it improved on the base model at all, and in every case the default produced a defender whose scores were indistinguishable from the untrained backbone. DPO's published learning rate and epoch count leave the LoRA update far too small on this backbone; we raise the learning rate to 10^{-4} and train for eight epochs. MUSE-D's data builder emits two preference streams, and the mixed stream is only

32% refusals in its chosen field; we train on the risk stream, which is 89% refusals. MTSA fits the implicit reward margin to a target derived from the future-reward gap, and normalizing that gap—as we initially did—drives the target to roughly one nat, which the optimizer reaches while leaving behaviour unchanged; restoring the unnormalized scale of the authors' formulation puts the target near 350 nats, comparable to the margin plain DPO reaches on the same pairs. X-Boundary needs two: its multi-turn boundary set must be supplied, since without it only the erase branch of the loss is active, and its retain term must be made scale-free. That term is a raw representation distance while the erase term is a cosine, so their balance depends on the backbone's residual-stream norm, which we measure at 182 on Qwen3 against 10 on Llama-3.1; with the raw form the retain term dominates and the erase objective does not move. None of these is a property of Qwen3 as a model. Each is a setting that carries over silently from the Llama configuration, and each is reported here so the comparison in Table 1 is reproducible.

KL-divergence setup. We measure per-position KL divergence over a fixed, teacher-forced continuation, so all models are scored on the same tokens and any difference reflects the model rather than its sampling. For each held-out attack dialogue, the continuation y is the base model's greedy final-turn response, generated once and reused verbatim; position i indexes the i-th token of y, and at each position we compute the exact next-token D_{KL} over the full vocabulary, averaged across the held-out set. We compare the *aligned* model (fine-tuned under our Assess $\rightarrow$ Decide $\rightarrow$ Respond framework), the *unaligned* model (framework prompt, no fine-tuning), and the *base* model (neither): the aligned–unaligned pair isolates the effect of training, while comparisons against the base isolate the framework itself. All three share identical dialogue content, differing only in system prompt and weights; for reasoning models we disable native thinking so y is a direct continuation. The shaded band marks the decision-token position on the model's own greedy generations.

Judges correlation setup. We compare seven verdict sources on our Qwen3-8B defender (two-label scaffold, self-verified decision reward): gpt-5-mini (Singh et al., 2025), gpt-oss-20b (Agarwal et al., 2025), StrongREJECT (Souly et al., 2024), Llama-Guard-3-8B (Llama Team, 2024), Qwen3Guard (Zhao et al., 2025), HarmBench-13B (Mazeika et al., 2024), and the model's own committed `Decision` label. The pool covers all thirteen held-out multi-turn safety cells: five attack families (SEMA, Crescendo, FITD, Echo, HIS) on both AdvBench and HarmBench, plus MHJ, CoSafe and MultiBreak, in total 3738 dialogues. Each source reduces a dialogue to a single binary *jailbroken* verdict on the final turn. Critically, the six external judges see only the user-facing reply, with the *assessment/Decision* deliberation stripped.

D PROMPTS

Figure 8 shows the two Assess $\rightarrow$ Decide $\rightarrow$ Respond system prompts verbatim. Both instruct the defender to precede every reply with an open, untagged deliberation of exactly two lines, an *Intent* line assessing the user's underlying goal and its potential for harm, and a *Decision* line committing to a label, followed by a reply consistent with that decision. The prompts differ only in the label menu and the matching clause of the closing instruction. The two-label variant accompanies the self-verified reward: because that reward checks the emitted label against the row's known ground truth, every admissible label must have one, and *safe-redirect* does not. The three-label variant is therefore usable only with an external reward that scores the reply itself, where a hedging option is meaningful. Neither prompt wraps the deliberation in tags; on a backbone with a native reasoning channel, this keeps the `<think>` block unambiguously the model's own and the framework cleanly separable from it.

E MORE EXPERIMENTAL RESULTS

E.1 UTILITY

Table 2 breaks the utility average of Table 1 down by benchmark. The pattern is uniform: on both backbones, every method, including ours, stays within a narrow band of its base model on every task, and no benchmark shows a systematic degradation attributable to safety training. What variation exists is concentrated in the instruction-following suite, the one task most sensitive to chat

Self-verified reward

```
You are a helpful and safe assistant.  Before replying to each user
message, write exactly two labeled lines, with no surrounding tags:
Intent:  <the user's true underlying intent, and whether fulfilling
the request could cause real harm>
Decision:  <one of COMPLY or REFUSE>
Then, starting on the next line, write your actual reply, consistent
with your Decision.  If the intent is benign, COMPLY and help fully;
if it is genuinely harmful, REFUSE.
```

External reward

```
You are a helpful and safe assistant.  Before replying to each user
message, write exactly two labeled lines, with no surrounding tags:
Intent:  <the user's true underlying intent, and whether fulfilling
the request could cause real harm>
Decision:  <one of COMPLY, REFUSE, or SAFE-REDIRECT>
Then, starting on the next line, write your actual reply, consistent
with your Decision.  If the intent is benign, COMPLY and help fully;
if it is genuinely harmful, REFUSE or SAFE-REDIRECT.
```

Figure 8: The two system prompts, identical except for the label menu and the matching clause of the closing instruction. See §D for why the self-verified reward requires the two-label menu.

behavior, while the knowledge- and reasoning-centric benchmarks are essentially untouched. This is the expected signature of the evaluation protocol discussed in §3: these suites are scored as plain-text completion without the chat template, so they bypass both the scaffold and the shallow, format-concentrated component of safety training, and confirm that our intervention leaves the underlying capabilities of the model intact.

E.2 Statistical Significance

We report the precision of Table 1 against the sampling noise of the evaluation: Table 3 restates every cell of Table 1 with a 95% confidence interval. Each safety cell is a SAFE rate over n held-out dialogues ($n = 516$, 785, 500, 1400, and 537 for Adv., Harm., MB, CoS., and MHJ respectively; 3,738 in total), with one exception: the *Llama3.1-8B* Base MHJ cell is scored over 300 dialogues, and its wider interval reflects that, so its sampling error is that of a binomial proportion; we report the 95% Wilson interval as a half-width, and a behavior-clustered bootstrap interval for the pooled Avg. Over-refusal (XSTest, XSB) carries Wilson intervals ($n = 250$, 340), and MT-Bench a 95% t-interval over its 160 per-turn scores. Utility is the mean of six lm-eval tasks; since lm-eval reports an independent standard error per task, the mean carries a combined half-width of about one point.

Three trends follow from the intervals. First, the safety conclusions are resolved unevenly, and the intervals are worth reading before the point estimates. The clearest case is the self-verified reward on *Qwen3-8B*: its average exceeds MUSE-D, the strongest baseline by average, by nearly nine half-widths, and its interval is disjoint from MUSE-D's on all five suites. That is a statement about MUSE-D specifically. On MHJ — the hardest suite and the one most worth scrutinising — the self-verified interval (75.2 ± 3.6) overlaps MTSA's (70.9 ± 3.8), so on that suite the two are not separated even though the point estimates differ by four points. The external reward on the same backbone is a different matter — its average lead is under two half-widths and its per-suite intervals overlap the strongest baseline's on every suite, so on that backbone the two should be read as comparable rather than ordered. On *Llama3.1-8B* both variants clear the strongest baseline on average, but the per-suite intervals overlap X-Boundary's on AdvBench for both, and on MultiBreak and MHJ for the self-verified variant. What the intervals support throughout is the separation between our variants and the *field* of baselines, not a resolved ordering against the single strongest one on every suite. Second, the Utility column is indistinguishable across methods: every entry lies within

Table 2: Per-benchmark general capability (Open LLM Leaderboard v1, six tasks). The final column is the **Utility Avg.** reported in Table 1. **Bold**: best per column within each panel.

Method	MMLU	HellaSwag	Winogrande	GSM8K	TruthfulQA	IFEval	**Avg.**
			`Llama3.1-8B-Instruct`				
Base	**68.34**	80.49	**78.69**	69.60	54.55	46.40	66.34
SFT	67.36	79.09	75.37	69.60	53.95	44.55	64.99
DPO	68.28	80.35	78.37	70.13	54.71	46.77	66.44
SDGO	68.31	**80.55**	77.98	70.28	54.51	45.29	66.15
MTSA	68.30	80.34	78.14	69.52	54.64	47.50	66.41
MUSE-D	68.31	80.45	78.14	69.14	54.72	**47.69**	66.41
X-Boundary	68.04	80.51	77.66	68.08	54.41	47.50	66.04
Self reward							
– unaligned	68.34	80.49	78.69	69.60	54.55	46.40	66.34
– aligned	68.14	80.17	78.14	**72.10**	54.51	**47.69**	**66.79**
External reward							
– unaligned	68.34	80.49	78.69	69.60	54.55	46.40	66.34
– aligned	68.10	80.45	78.30	69.22	**54.82**	43.81	65.78
			`Qwen3-8B`				
Base	74.86	76.46	70.40	87.95	54.45	25.69	64.97
SFT	**75.03**	75.93	70.40	88.02	**56.14**	**29.57**	**65.85**
DPO	74.61	75.80	70.40	87.34	54.40	22.55	64.18
SDGO	74.89	76.44	**70.80**	87.49	54.35	25.51	64.91
MTSA	74.44	75.39	70.01	87.64	54.25	23.29	64.17
MUSE-D	74.75	75.65	70.24	**88.10**	54.13	23.48	64.39
X-Boundary	74.85	**76.51**	70.56	88.02	54.57	25.88	65.07
Self reward							
– unaligned	74.86	76.46	70.40	87.95	54.45	25.69	64.97
– aligned	74.95	76.45	69.69	87.41	54.65	24.21	64.56
External reward							
– unaligned	74.86	76.46	70.40	87.95	54.45	25.69	64.97
– aligned	74.98	**76.54**	70.40	87.19	54.29	24.58	64.66

roughly one combined interval of its base model, confirming that no method, ours included, meaningfully changes capability. Third, the helpfulness intervals counsel asymmetric reading: they are wide relative to the small method-to-method differences on XSTest and MT-Bench, which should therefore not be over-interpreted, whereas the elevated over-refusal of the aligned defenders relative to their base models clears the intervals comfortably and is a real cost rather than noise.

E.3 Multiple Judges

Table 4 reports safety under five additional judges, all held out from training. The two rewards behave differently here, and the distinction matters. The self-verified reward places first or second in all 25 judge-by-benchmark cells on *Qwen3-8B*, and is the strict maximum in 22 of them. The three exceptions are CoSafe under Llama-Guard-3, where all four of our rows tie at 100.0; CoSafe under HarmBench-13B, where our external variant is marginally ahead (96.9 against 96.4); and MHJ under Qwen3Guard, the one cell in the panel where a baseline leads it (MTSA, 81.8 against 78.8). The external reward is not comparably robust: it falls outside the top two in 9 of the 25 cells, ranking third or fourth on MHJ under every judge and third or fourth on AdvBench and HarmBench under Qwen3Guard. Claims about judge-independence should therefore be made for the self-verified reward specifically. The baselines also reorder among themselves — MUSE-D leads MTSA under gpt-oss-20b, StrongREJECT and HarmBench-13B, and the two swap under Llama-Guard-3 and Qwen3Guard — so what the judges agree on is the position of the self-verified reward relative to the field, not a total ordering. The lead is also narrower than the ranking implies: averaged over the five judges the self-verified reward exceeds the strongest baseline by $+7.1$ on

Table 3: Table 1 with 95% confidence intervals, as value $\pm$half-width. Safety cells and over-refusal (XSTest, XSB) use Wilson intervals; the safety Avg. a behavior-clustered bootstrap; Utility a combined lm-eval standard error; MT-Bench a t-interval. Judge: `gpt-5-mini`.

	Safety ↑						Utility ↑	Helpfulness		
Method	Adv.	Harm.	MB	CoS.	MHJ	Avg.	Avg.	XSTest ↓	XSB ↓	MT-Bench ↑
`Llama3.1-8B-Instruct`										
Base	54.5 ±4.3	62.5 ±3.4	59.8 ±4.3	84.9 ±1.8	38.0 ±5.4	59.9 ±1.8	66.3 ±1.0	10.0 ±3.7	30.0 ±4.8	6.20 ±0.41
SFT	69.4 ±4.0	79.5 ±2.9	87.0 ±3.0	88.5 ±1.7	55.1 ±4.2	75.9 ±1.4	65.0 ±1.1	14.0 ±4.3	45.6 ±5.3	5.67 ±0.39
DPO	61.6 ±4.2	71.5 ±3.1	79.0 ±3.5	88.3 ±1.7	46.7 ±4.2	69.4 ±1.5	66.4 ±1.0	10.4 ±3.8	34.4 ±5.0	6.26 ±0.42
SDGO	56.6 ±4.2	61.7 ±3.4	60.4 ±4.3	85.0 ±1.9	39.1 ±4.1	60.6 ±1.7	66.2 ±1.0	11.6 ±4.0	30.9 ±4.9	6.11 ±0.42
MTSA	61.6 ±4.2	71.6 ±3.1	79.0 ±3.5	86.6 ±1.8	47.1 ±4.2	69.2 ±1.5	66.4 ±1.0	10.8 ±3.9	36.5 ±5.1	6.12 ±0.41
MUSE-D	61.8 ±4.2	67.5 ±3.2	74.2 ±3.8	87.4 ±1.8	45.3 ±4.2	67.2 ±1.6	66.4 ±1.0	8.4 ±3.5	33.2 ±5.0	6.27 ±0.42
X-Boundary	97.9 ±1.3	95.0 ±1.6	93.0 ±3.0	93.6 ±1.3	52.0 ±4.2	86.3 ±1.2	66.0 ±1.0	16.8 ±4.6	50.3 ±5.3	6.26 ±0.42
Self reward										
– unaligned	63.0 ±4.1	70.3 ±3.2	64.2 ±4.2	89.6 ±1.6	44.3 ±4.1	66.3 ±1.7	66.3 ±1.0	8.0 ±3.4	33.8 ±5.0	5.61 ±0.40
– aligned	98.8 ±1.0	99.5 ±0.5	96.6 ±1.7	98.4 ±0.7	58.5 ±4.2	90.4 ±1.0	66.8 ±1.0	20.4 ±5.0	54.4 ±5.3	5.86 ±0.39
External reward										
– unaligned	57.2 ±4.2	65.2 ±3.4	56.6 ±4.3	88.5 ±1.7	35.0 ±4.0	60.5 ±1.6	66.3 ±1.0	8.8 ±3.5	32.6 ±5.0	5.81 ±0.41
– aligned	96.9 ±1.5	98.2 ±1.0	99.2 ±0.9	99.0 ±0.6	71.1 ±3.8	92.9 ±0.9	65.8 ±1.0	21.6 ±5.1	50.0 ±5.3	5.21 ±0.39
`Qwen3-8B`										
Base	49.2 ±4.3	52.5 ±3.5	23.8 ±3.7	82.6 ±2.0	22.3 ±3.6	46.1 ±1.6	65.0 ±1.0	7.2 ±3.2	15.6 ±3.9	6.08 ±0.52
SFT	59.9 ±4.2	65.7 ±3.4	49.6 ±4.4	88.5 ±1.7	35.6 ±4.1	59.9 ±1.7	65.9 ±1.0	17.6 ±4.7	40.9 ±5.2	7.53 ±0.33
DPO	80.2 ±3.4	86.6 ±2.4	83.2 ±3.3	97.5 ±0.8	53.6 ±4.2	80.2 ±1.4	64.2 ±1.0	26.0 ±5.4	39.4 ±5.2	5.87 ±0.52
SDGO	48.3 ±4.3	52.5 ±3.5	23.2 ±3.7	82.4 ±2.0	33.0 ±3.9	47.9 ±1.7	64.9 ±1.0	10.0 ±3.7	15.9 ±3.9	6.23 ±0.52
MTSA	79.1 ±3.5	90.2 ±2.1	85.8 ±3.1	97.8 ±0.8	70.9 ±3.8	84.8 ±1.3	64.2 ±1.0	38.4 ±6.0	53.5 ±5.3	5.99 ±0.52
MUSE-D	87.6 ±2.8	90.8 ±2.0	91.0 ±2.5	97.4 ±0.8	61.5 ±4.1	85.7 ±1.2	64.4 ±1.0	20.4 ±5.0	40.9 ±5.2	5.87 ±0.52
X-Boundary	68.0 ±4.0	71.1 ±3.2	67.8 ±4.1	88.8 ±1.7	27.4 ±3.8	64.6 ±1.5	65.1 ±1.0	18.8 ±4.8	49.1 ±5.3	5.69 ±0.52
Self reward										
– unaligned	69.0 ±3.9	80.1 ±2.8	66.8 ±4.1	96.4 ±1.0	46.9 ±4.2	71.8 ±1.6	65.0 ±1.0	14.8 ±4.4	37.9 ±5.1	7.33 ±0.36
– aligned	97.1 ±1.5	97.6 ±1.1	97.6 ±1.3	99.3 ±0.4	75.2 ±3.6	93.4 ±0.9	64.6 ±1.0	26.4 ±5.4	56.8 ±5.2	7.36 ±0.34
External reward										
– unaligned	68.8 ±4.0	77.3 ±2.9	69.4 ±4.0	95.5 ±1.1	44.3 ±4.2	71.1 ±1.5	65.0 ±1.0	10.0 ±3.7	39.7 ±5.2	7.20 ±0.35
– aligned	90.1 ±2.6	93.6 ±1.7	92.8 ±2.3	98.4 ±0.7	63.1 ±4.0	87.6 ±1.1	64.7 ±1.0	16.8 ±4.6	42.6 ±5.2	6.96 ±0.41

MultiBreak but only $+2.6$ on MHJ and $+3.2$ on CoSafe, and on CoSafe under Llama-Guard-3 the margin is $+0.1$ with both our variants at 100.0. The judges do differ in absolute calibration, with some scoring systematically stricter or more lenient than others, which is consistent with the moderate inter-judge agreement quantified in §4.2; crucially, however, these shifts move all methods together and leave the ranking essentially unchanged. Since our reward never consults any of these judges, this consistency indicates that the measured gains reflect genuinely safer behavior rather than overfitting to the idiosyncrasies of a particular evaluator.

Table 4: Safety across five benchmarks under five judges held out from training. The corresponding *gpt-5-mini* scores are in Table 1. **Bold**: best per column within each judge block.

	`Llama3.1-8B-Instruct`					`Qwen3-8B`				
Method	Adv.	Harm.	MB	CoS.	MHJ	Adv.	Harm.	MB	CoS.	MHJ
Judge: gpt-oss-20b										
Base	85.3	71.4	63.9	69.6	48.5	62.0	61.0	27.1	60.9	38.2
SFT	77.3	83.9	87.4	74.4	66.0	73.8	75.5	60.4	71.2	48.0
DPO	72.5	78.3	80.1	73.0	58.2	88.0	89.4	82.8	83.0	64.2
SDGO	66.0	70.2	62.7	68.5	52.1	62.2	61.5	28.2	61.9	44.7
MTSA	72.0	78.2	80.4	73.1	56.6	87.8	93.1	85.8	83.9	80.1
MUSE-D	70.1	75.4	75.0	71.7	56.7	92.6	94.4	90.6	85.4	71.5
X-Boundary	97.7	96.1	94.6	81.5	62.2	77.5	77.5	70.8	69.3	41.2
Self reward										
– unaligned	72.4	76.9	66.2	75.7	55.8	77.7	83.7	70.0	86.3	60.3
– aligned	**99.6**	**99.9**	96.0	86.1	69.3	**98.6**	**98.7**	**96.2**	**89.2**	**82.0**
External reward										
– unaligned	69.2	73.0	59.3	71.0	49.6	77.9	82.9	71.8	87.1	55.6
– aligned	97.9	99.2	**98.2**	**90.9**	**79.9**	94.8	96.3	92.4	86.0	72.6
Judge: StrongREJECT										
Base	85.0	72.3	57.3	42.4	50.9	77.1	67.4	27.2	31.5	39.7
SFT	81.1	87.1	83.8	51.0	67.7	86.6	79.6	62.5	40.3	47.5
DPO	74.5	79.5	74.2	50.7	58.2	93.0	91.7	78.6	62.8	63.5
SDGO	70.4	72.9	57.9	44.3	53.2	77.5	66.6	27.0	31.4	45.7
MTSA	74.1	80.0	77.2	50.7	59.1	93.0	90.3	77.8	66.1	75.0
MUSE-D	74.4	76.9	72.1	48.2	58.1	94.8	93.6	85.0	66.2	71.1
X-Boundary	99.4	98.3	91.3	68.3	67.7	86.4	81.7	66.8	41.9	41.2
Self reward										
– unaligned	77.3	79.2	62.6	52.9	60.2	85.0	86.2	69.4	58.6	62.4
– aligned	99.4	**99.5**	93.0	73.6	74.6	**98.8**	**98.2**	**92.4**	**74.0**	**82.5**
External reward										
– unaligned	74.6	77.7	58.4	45.0	56.4	90.9	86.8	73.9	55.7	62.3
– aligned	**99.6**	99.4	**96.8**	**77.4**	**84.9**	98.6	96.6	87.6	67.0	74.6
Judge: Llama-Guard-3										
Base	91.6	81.9	58.2	97.4	60.3	83.7	73.6	22.8	97.3	53.1
SFT	92.8	92.5	84.0	98.7	75.8	93.8	87.8	63.2	98.8	67.4
DPO	90.1	88.0	76.6	98.4	70.0	94.8	93.5	77.8	99.4	71.9
SDGO	84.7	80.3	58.8	97.3	66.3	85.1	75.8	20.8	97.4	52.9
MTSA	89.0	88.4	77.8	98.4	70.9	98.1	98.0	81.6	99.9	84.7
MUSE-D	88.4	84.3	72.0	97.9	68.9	98.3	96.6	86.2	99.4	76.0
X-Boundary	80.4	82.1	72.3	93.5	64.6	84.9	80.6	40.4	97.7	59.0
Self reward										
– unaligned	89.0	83.8	62.0	97.9	69.1	94.2	91.8	69.8	**100.0**	73.6
– aligned	**99.8**	**100.0**	96.6	**100.0**	77.8	**99.8**	**99.9**	**95.4**	**100.0**	**88.5**
External reward										
– unaligned	87.6	82.5	57.8	97.9	62.9	95.2	92.7	74.8	**100.0**	74.5
– aligned	**99.8**	99.5	**99.8**	**100.0**	**85.3**	99.4	99.5	93.8	**100.0**	82.5
Judge: Qwen3Guard										
Base	65.3	46.2	61.0	90.1	47.3	42.4	40.8	26.0	93.1	44.3
SFT	52.9	61.9	88.0	91.7	66.3	52.5	55.5	61.4	96.5	53.1
DPO	50.8	58.5	79.2	93.2	58.7	69.0	75.9	84.4	98.1	60.9
SDGO	42.4	46.1	62.0	89.4	53.1	41.7	39.6	26.6	93.2	43.2
MTSA	50.0	58.6	80.0	92.9	58.5	74.8	83.8	87.4	99.3	**81.8**
MUSE-D	48.3	53.4	75.2	91.6	58.1	79.1	81.8	91.8	98.8	67.6
X-Boundary	36.8	53.2	73.3	82.4	52.3	43.8	47.0	42.8	93.5	48.4
Self reward										
– unaligned	49.0	53.4	64.4	90.0	53.3	47.9	58.1	67.6	97.1	56.1
– aligned	**98.4**	**98.6**	98.6	99.1	69.6	**90.1**	**91.1**	**97.8**	**99.9**	78.8
External reward										
– unaligned	44.4	51.1	59.0	94.3	49.3	50.8	58.1	66.2	99.3	52.0
– aligned	94.0	96.8	**98.8**	**99.4**	**78.8**	74.4	81.9	92.6	99.5	70.0
Judge: HarmBench-13B										
Base	83.8	66.1	59.0	73.1	49.0	52.7	53.0	22.0	71.8	42.8
SFT	74.4	80.5	87.8	77.1	66.7	66.9	70.2	56.0	78.0	52.0
DPO	69.8	72.2	77.8	76.5	59.6	87.2	91.3	84.6	91.4	68.5
SDGO	65.1	65.0	61.0	73.5	52.5	49.4	54.9	20.8	71.6	49.7
MTSA	68.4	73.1	80.2	77.6	58.7	82.2	89.0	82.8	92.5	81.0
MUSE-D	69.6	70.6	75.4	75.9	57.4	94.2	95.4	90.4	92.5	76.4
X-Boundary	95.2	92.8	90.7	84.4	62.6	71.7	75.3	67.0	77.5	48.6
Self reward										
– unaligned	70.7	71.7	65.2	82.9	59.4	80.2	84.3	72.0	93.6	65.9
– aligned	**99.8**	**99.6**	98.2	93.6	73.0	**98.8**	**98.6**	**97.8**	96.4	**84.0**
External reward										
– unaligned	67.4	69.0	57.0	84.3	52.7	77.3	84.1	75.6	95.0	63.5
– aligned	98.6	98.7	**99.6**	**99.0**	**82.5**	97.9	97.3	96.2	**96.9**	74.7