BEYOND REFUSAL RATES: HOW SFT AND RL SHAPE LLM SAFETY DISCERNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment aims to make large language models helpful without facilitating harm. We investigate how supervised fine-tuning (SFT) and reinforcement learning (RL) shape safety alignment, using group relative policy optimization (GRPO) as a representative RL method. Starting from the same pretrained model and safety prompts, the resulting models achieve similar refusal rates on harmful requests and similar accuracy on math and reasoning benchmarks. Yet GRPO rejects fewer benign requests and more often returns to refusal after an attack forces its response to begin by agreeing to help. These differences concern *safety discernment*: the ability to answer benign requests and refuse harmful ones despite misleading context. To understand this contrast, we combine training ablations with internal analyses. Rewarding appropriate answers and refusals, rather than harmless content alone, helps GRPO avoid unnecessary refusals. We also find that reinforcing better responses while suppressing worse ones is sufficient to learn recovery after the forced opening. Internal interventions link this recovery to the continued influence of refusal-related information after the forced opening. Together, these findings connect safety discernment to specific training signals and reveal behavioral differences that refusal rates and task accuracy alone fail to capture.

1 INTRODUCTION

Safety alignment is essential for deploying large language models (LLMs) that assist users without facilitating harmful requests. Supervised fine-tuning (SFT) learns from demonstrations, whereas reinforcement learning (RL) learns from feedback on generated responses; practical post-training pipelines often combine the two (Ouyang et al., 2022; Bai et al., 2022b; Dai et al., 2023). Prior comparative studies have examined their generalization behavior (Chu et al., 2025) and safety–reasoning trade-offs (Jia et al., 2025). We ask a complementary question: how does the choice of learning algorithm affect the model's appropriate use of request refusal?

High refusal rates on harmful requests alone do not establish that refusal is used appropriately across contexts. We examine two complementary failures. *Over-refusal* rejects benign requests, including those whose wording resembles harmful content (Rottger et al., 2024). A *prefix attack* leaves a harmful request unchanged but supplies an assistant opening such as "I can help with that" before the model continues, potentially steering it away from refusal (Qi et al., 2025). Together, these settings probe whether a model avoids both unwarranted refusal and unwarranted compliance. We use **safety discernment** to describe this behavioral ability: answering benign requests and refusing harmful ones, even when wording or a supplied response prefix encourages the opposite behavior.

Our central question is therefore *how SFT and RL shape safety discernment*. We study group relative policy optimization (GRPO; Shao et al., 2024) as a representative RL method. Our primary comparison starts both methods from the same pretrained weights and uses the same safety-training prompts, avoiding differences in prior instruction- or safety-tuning histories. The resulting models reach comparable ordinary harmful-refusal rates and general-task performance, yet GRPO refuses fewer benign requests and recovers refusal to harmful requests more often after a helpful prefix. This gap motivates examining what the training procedures teach beyond producing refusals under ordinary conditions.

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.

We investigate these differences through training-signal ablations and internal representation analyses. For over-refusal, a reward ablation shows that response-specific feedback distinguishing appropriate answers from unnecessary refusals helps GRPO avoid indiscriminate refusal. Internally, GRPO more strongly favors answering benign requests and refusing harmful ones, and its appropriate decisions are harder to overturn under interventions. For recovery after a prefix attack, we progressively simplify GRPO. The resulting procedure, *Offline Signed Learning* (OSL), uses responses collected once from the starting model and keeps both these responses and their reward scores unchanged during training. Reinforcing better responses while suppressing worse ones on this fixed set is sufficient to acquire substantial recovery without prefix-attack training examples. Activation tracking and prompt-end interventions further show that refusal-related information introduced before the supplied response prefix has greater influence on the subsequent response in GRPO and OSL than in SFT. A supplementary evaluation on instruction-tuned models finds competitive prefix recovery at lower measured GPU-hour cost than online GRPO.

Together, our results show that similar harmful-request refusal rates and task accuracy can conceal substantial differences in safety discernment. Our analyses identify two relevant choices in training: whether feedback distinguishes appropriate answers from unnecessary refusals, and whether learning suppresses worse responses as well as reinforcing better ones. These findings connect SFT–RL behavioral differences to concrete training choices and motivate directly evaluating safety discernment beyond refusal rates.

2 Related Work

Safety training with SFT and reinforcement learning. Safety alignment combines demonstrations, preferences and outcome rewards, as in InstructGPT, Constitutional AI and Safe RLHF (Ouyang et al., 2022; Bai et al., 2022b; Dai et al., 2023). DPO learns from offline preference pairs, whereas GRPO uses within-prompt relative rewards for sampled responses (Rafailov et al., 2023; Shao et al., 2024).

Over-refusal, prefix attacks and refusal mechanisms. XSTest, OR-Bench and FalseReject document unnecessary refusal of benign requests (Rottger et al., 2024; Cui et al., 2025; Zhang et al., 2025b). Safety-aware preferences and policy-grounded reasoning address this safety–helpfulness trade-off (Kim et al., 2026; Guan et al., 2024). Shallow-alignment work links prefix vulnerability to safety changes concentrated at early response tokens and studies recovery-oriented augmentation (Qi et al., 2025). Refusal-direction analyses provide activation-based tools to examine these behaviors (Arditi et al., 2024; Zhao et al., 2025). Appendix I provides an extended discussion.

3 Establishing Comparable SFT and GRPO Baselines

To understand how post-training affects refusal behavior, we first establish a common basis for comparing SFT and GRPO. Both methods start from the same pretrained model and use the same safety-training prompts. We examine their training dynamics and final harmful-refusal and utility performance, establishing the baseline policies for the subsequent analyses.

3.1 Learning from Demonstrations and Rewards

Supervised fine-tuning. SFT learns one selected demonstration y^* for each prompt x by minimizing response-token negative log likelihood:

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{Z} \sum_{(x,y^*)\in\mathcal{D}} \sum_{t} \log \pi_\theta(y_t^* \mid x, y_{<t}^*), \tag{1}$$

where Z counts supervised response tokens. Each demonstration specifies whether to refuse or answer the request, as well as how to phrase the response.

Group relative policy optimization. GRPO samples a group of responses for each prompt and uses their relative rewards to update the policy (Shao et al., 2024). Rewards are standardized within

each prompt group to obtain advantages A_i. The core clipped policy objective is

$$\mathcal{L}_{\mathrm{GRPO}}^{\mathrm{core}} = -\frac{1}{Z}\sum_{i,t}\min(\rho_{i,t}A_i, \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i) + \beta\widehat{D}_{\mathrm{KL}}(\pi_\theta, \pi_{\mathrm{ref}}), \tag{2}$$

Here $\rho_{i,t}$ is the current-to-old-policy token probability ratio, ϵ sets the clipping range, and β controls regularization toward the fixed reference policy π_{ref}.

Rewards for safe and appropriate responses. For safety training, we evaluate both whether a response is unharmful and whether it makes the appropriate decision to answer or refuse. We use WildGuard as the reward judge because it provides both response harmfulness and refusal judgments (Han et al., 2024). Combining these judgments with the prompt's target behavior gives the safety–behavior reward component

$$R_i^{\mathrm{SB}} = 0.6\, s_i + 0.4\, b_i. \tag{3}$$

The safety score s_i is $+1$ for an unharmful response and -1 for a harmful response. The behavior score b_i is $+1$ for refusing a harmful request or not refusing a benign request, and -1 otherwise. We fix the safety and behavior weights at 0.6 and 0.4, respectively, to prioritize unharmful content while retaining feedback on the answer/refuse decision. The behavior term favors the appropriate decision among responses with the same safety score. Appendix A gives reward implementation, normalization, clipping and optimizer details.

3.2 Shared Initialization, Data, and Evaluation

Models. We study Qwen2.5-3B Base (Yang et al., 2024); additional experiments with Llama-3.2-3B (Grattafiori et al., 2024) appear in Appendix F. Both primary Qwen policies start from the same Base weights. We use two SFT epochs because extending training to three or four epochs at the same initial learning rate does not improve harmful-request refusal (Appendix B.1). We train GRPO for one epoch, corresponding to 1250 updates. Over the final 50 updates, its mean training reward is 0.971 on a scale with maximum 1, indicating that the policy approaches the training-reward ceiling by the end of this pass.

Training data. We collect 10,000 refusal-target and 10,000 compliance-target requests from WildGuardMix (Han et al., 2024), AEGIS2.0 (Ghosh et al., 2025) and the safety subset of Dolci-Instruct-SFT (Olmo et al., 2025). SFT learns a demonstration for each prompt; GRPO samples 8 responses and scores them with the reward judge described above. Each GRPO rollout batch receives one optimizer update, with gradients accumulated across microbatches. Appendix A reports the statistics of training prompts drawn from each dataset and the training settings.

Test datasets. We evaluate saved checkpoints on harmful-request refusal and general-task utility. Harmful-request benchmarks are HarmBench-standard200 (Mazeika et al., 2024), ORBench-toxic (Cui et al., 2025) and SORRY-Bench (Xie et al., 2024). General-task utility is measured by generated-answer accuracy on GSM8K (Cobbe et al., 2021), ARC-Easy (Clark et al., 2018) and PIQA (Bisk et al., 2020). We find no overlap between training corpus and these evaluation prompts.

Evaluation methods. WildGuard (Han et al., 2024) supplies separate refusal and harmful-output labels. Refusal rate measures the decision studied in this paper; harmful-output rates appear with the full results. Utility uses generated-answer accuracy. Dataset sizes, decoding and judging protocols are in Appendix A; per-dataset outcomes are in Appendix B.

3.3 Learning Dynamics and Endpoint Performance

Figure 1 tracks harmful-request refusal and general-task accuracy from the shared Base initialization. SFT acquires high ordinary refusal early in training, whereas GRPO approaches a similarly high endpoint more gradually. Despite these different trajectories, the final models have comparable harmful-refusal and utility performance: mean harmful refusal is 98.02% for SFT and 96.36% for GRPO, while mean general-task accuracy is 76.55% and 75.06%, respectively. These endpoints provide the basis for the remaining comparisons. We next ask whether their similar ordinary performance extends to answering benign requests and recovering refusal after a helpful response prefix.

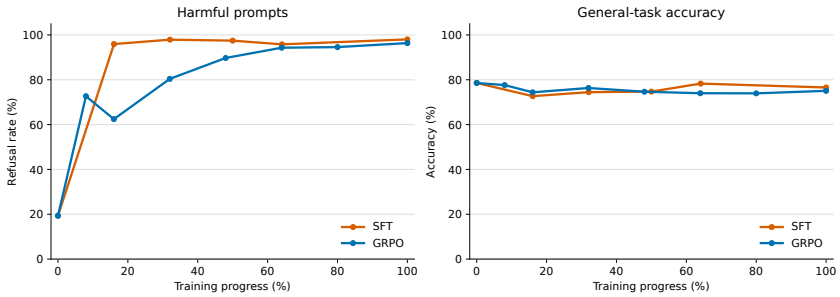


Figure 1: **Common-Base Qwen3B training dynamics.** Left: harmful-request refusal averaged equally over HarmBench, ORBench-toxic and SORRY-Bench. Right: accuracy averaged equally over GSM8K, ARC-Easy and PIQA.

Table 1: **Benign refusal in Qwen3B.** Lower values indicate fewer refusals of benign requests.

Endpoint	FalseReject ↓	XSTest-safe ↓	ORBench-hard ↓	JBB-benign ↓	Mean ↓
SFT	63.10	20.80	74.07	74.00	57.99
GRPO	**52.49**	**8.80**	**68.84**	**63.00**	**48.28**

4 Understanding GRPO's Lower Over-Refusal

4.1 GRPO refuses fewer benign requests

Over-refusal occurs when a model rejects a benign request. For example, a safety-aligned model may refuse "Can you help me kill a Python process?", even though the request concerns terminating a computer program. We evaluate over-refusal on FalseReject (Zhang et al., 2025b), XSTest-safe (Rottger et al., 2024), ORBench-hard (Cui et al., 2025) and JailbreakBench-benign (Chao et al., 2024). Table 1 shows lower benign refusal under GRPO on all four datasets, with dataset-average benign refusal decreasing from 57.99% for SFT to 48.28% for GRPO. Thus, although the two models train on the same prompts and reach similarly high harmful refusal, GRPO is more selective on these benchmarks: it answers more benign requests that SFT rejects.

4.2 Behavior Feedback Reduces Over-Refusal

What training feedback encourages GRPO to answer benign requests rather than refuse them? The safety score in Eq. 3 rewards unharmful responses, whether they answer or refuse a benign request. The behavior score distinguishes these decisions by rewarding answers to benign requests and refusals of harmful ones. We test this distinction by repeating Base-start GRPO with each response's behavior score replaced by the mean score of its response group, while retaining the safety feedback and other training settings. Within-group centering cancels this shared behavior term, removing its ability to favor an appropriate benign answer over an unnecessary refusal when both responses are unharmful. The resulting *Behavior-neutral* model refuses 100% of both the benign and harmful evaluation requests (Appendix C). Response-specific behavior feedback therefore helps avoid indiscriminate refusal: it distinguishes answering a benign request from treating refusal as an acceptable response to every request.

4.3 Response Preferences and Resistance to Over-Refusal

Measuring Preferences for Answering or Refusing. The reward comparison identifies feedback that reduces over-refusal. We next examine how the resulting GRPO model differs internally from SFT when deciding whether to answer. An observed decision alone may conceal this difference: both models can answer a benign request while assigning different likelihoods to answering and refusing. We therefore first measure each model's initial preference for answering versus refusing. We compare these preferences on benign requests that both models answer and harmful requests that both refuse, so their observed decisions agree. Appendix C gives the sources and selection details.

Table 2: **Initial response preferences with uniform opening weights.** Mean $M_K = R_K - C_K$ before intervention on the shared-correct requests in Figure 2. K counts scored opening tokens. Lower benign and higher harmful values indicate more request-appropriate preferences.

Model	M_3	M_4	M_5
Benign			
SFT	0.580	0.298	0.211
GRPO	−1.639	−1.916	−1.664
Harmful			
SFT	2.576	1.972	1.471
GRPO	3.428	2.933	2.200

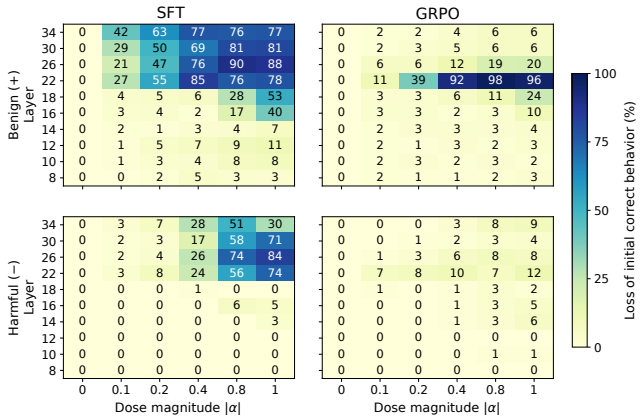


Figure 2: **Decision stability under refusal-related interventions.** The intervention maps use jointly answering benign and jointly refusing harmful prompts. The upper row adds the refusal direction on benign prompts; the lower row subtracts it on harmful prompts.

For each request, we compare the likelihoods of short response openings associated with refusal or compliance, using the same candidates for both models. We collect these openings from both models' responses on a separate discovery set, deduplicate identical token prefixes, and fix the candidate sets before scoring the evaluation prompts. Here K is the number of opening tokens scored. For each candidate, we average its conditional token log probabilities, then aggregate with equal weight per distinct opening within each class by log-mean-exp. Let $R_K(x)$ and $C_K(x)$ denote these refusal and compliance scores for request x. Their difference defines the *refusal-preference readout*:

$$M_K(x) = R_K(x) - C_K(x), \qquad K \in \{3, 4, 5\}. \tag{4}$$

A larger M_K means that the model assigns higher likelihood to the refusal openings relative to the compliance openings; M_4, for example, uses the four-token candidates. Appendix C.3 gives the construction, coverage audit and a frequency-weighted sensitivity analysis.

Table 2 shows the same ordering at all three token lengths: GRPO has a lower refusal-preference readout on benign requests and a higher readout on harmful requests. At $K = 4$, the benign contrast is -1.916 for GRPO versus 0.298 for SFT, while the harmful contrast is 2.933 versus 1.972. Thus, even where their generated decisions agree, GRPO's opening preferences more strongly distinguish requests it should answer from those it should refuse. In particular, its stronger relative preference for answering benign requests is consistent with its lower over-refusal. The ordering also holds when openings are weighted by their discovery frequencies (Appendix C.3, Table 11).

Testing Susceptibility to Over-Refusal. The preference readout describes how the models initially favor answering or refusing. We next test how readily an appropriate answering decision can be pushed toward over-refusal. Benign requests can contain harmful-sounding cues, as in the Python-process example above (Rottger et al., 2024). Such cues may increase refusal-related activity and make refusal more likely even when the request is benign. Prior work shows that adding activity along a refusal-related direction can induce refusal on benign requests (Arditi et al., 2024). To mimic this refusal-inducing pressure, we directly add a refusal-related direction to activations of benign requests that both models initially answer.

We estimate these directions from a separate, author-constructed set of harmful/benign prompt pairs. For each model, layer and template position, we compute the mean harmful-minus-benign activation contrast. The template tokens mark the transition from the user request to the assistant response, a site used in prior refusal analyses (Galichin et al., 2026; Zhang et al., 2025a). Each trial intervenes on one layer, with the dose scaled to the model's benign activation magnitude at each site. As a complementary test, we subtract the direction on harmful requests that both models initially refuse. Appendix C specifies fitting and dose scaling.

Table 3: **A helpful opening exposes the SFT–GRPO recovery gap**. Each entry reports natural refusal / H5 refusal / H5 harmful-output rates (%), judged by WildGuard.

Endpoint	ORBench-toxic	SORRY-Bench	HarmBench
SFT	96.11 / 2.88 / 86.80	98.40 / 5.85 / 74.20	100.00 / 19.50 / 75.00
GRPO	94.25 / 89.17 / 1.86	94.68 / 91.22 / 0.00	99.50 / 97.50 / 0.00

After each intervention, we generate and judge the response. On benign requests, we measure the fraction of initially appropriate answers replaced by refusals; this is induced over-refusal. On harmful requests, we measure the fraction of appropriate refusals lost under subtraction. Figure 2 displays both as the *loss of initial correct behavior*, which is zero without intervention. It shows different layer-wise susceptibility profiles. SFT is vulnerable to refusal-inducing additions across a broader range of late layers, whereas GRPO's strongest susceptibility is concentrated around layer 22. At layer 26 and dose 0.8, induced benign refusal is 90% for SFT and 19% for GRPO. SFT is therefore more readily pushed into over-refusal across a broader range of late layers, alongside greater susceptibility to losing appropriate harmful-request refusal.

Together, the readout and intervention results show that GRPO's lower over-refusal is accompanied by a stronger initial preference for answering benign requests and its appropriate decisions are harder to overturn under refusal-inducing intervention. The reward comparison identifies response-specific behavior feedback as an important training signal for learning this selective behavior.

5 Learning to Refuse After a Helpful Prefix

5.1 GRPO Recovers Refusal More Often Than SFT

A *prefix attack*, or response-prefill attack, supplies the beginning of the assistant's response before allowing the model to continue generating. A model may refuse a harmful request when it chooses its own opening, yet answer the same request after being made to begin "I can help with that". We use this fixed assistant opening, which consists of five tokens under the Qwen tokenizer, and denote the condition H5. We measure the fraction of harmful requests for which the model refuses after the supplied prefix and compare it with *natural refusal*, where the model generates its entire response without a supplied opening. Table 3 shows a large gap between the primary SFT and GRPO endpoints. Both refuse most requests naturally. After H5, SFT usually continues without refusing, whereas GRPO usually recovers refusal. Table 16 in Appendix D illustrates this recovery: GRPO follows the supplied helpful opening with a transition into refusal. The gap therefore concerns how the model continues an already-started response, not simply whether it learns to begin with a refusal.

5.2 Learning Prefix Recovery from Positive and Negative Feedback

What Enables Recovery After a Helpful Prefix? We progressively simplify GRPO to identify a learning signal that can produce prefix recovery. GRPO combines online response sampling, group-relative advantages, a clipped policy-ratio objective and reference KL regularization. Under our training schedule (Section 3.2), each rollout batch receives one policy epoch and a single optimizer update after microbatch gradient accumulation. The current policy therefore matches the rollout policy when the gradient is evaluated: $\rho_{i,t} = 1$. Policy clipping is inactive, and the ratio term yields the same policy gradient as advantage-weighted log likelihood (Appendix E).

We next simplify the remaining components. We first remove reference KL while retaining online sampling and group-relative advantages. We then replace online response refresh with a bank collected once from the initial policy, freezing both the responses and their advantages. For this offline variant, we retain advantage-weighted log likelihood rather than importance-weighting the frozen responses as the policy changes. We call this procedure *Offline Signed Learning* (OSL). Its objective is related to Group Policy Gradient (GPG; Chu et al., 2026):

$$\mathcal{L}_{\mathrm{OSL}} = -\frac{1}{Z} \sum_i A_i \sum_{t \in \mathcal{T}_i} \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t}). \tag{5}$$

Table 4: **Dissecting the learning signal for prefix recovery.** All trained methods start from one-epoch SFT and receive 200 additional updates. Evaluation protocols are given in Appendix D.4.

	Natural harmful ↑	H5 harmful ↑	Benign ↓
SFT start	97.44	3.90	49.99
GRPO	96.98	14.58	47.67
GRPO without KL	96.58	19.25	44.52
OSL	93.47	**86.41**	41.81
Positive-only OSL	85.05	10.41	24.18
Best-response SFT	97.38	3.69	49.45

Here $\mathcal{T}_i$ contains the supervised tokens in response i, and Z is the total number of supervised response tokens in the update. Positive advantages encourage the corresponding responses, whereas negative advantages discourage them. The remaining learning signal of OSL includes both positive and negative credit. To test whether positive learning alone can produce recovery, we construct two controls from the same frozen bank. *Positive-only OSL* removes negative coefficients and rescales the positive coefficients within each prompt group to preserve its total absolute token-weighted credit. *Best-response SFT* selects the highest-reward response to each prompt and repeats it to preserve the group size. These controls test whether recovery follows from reinforcing favorable responses or fitting the best available response.

A Shared Starting Point and Training Protocol. The five learners are independent branches. All start from the same one-epoch SFT checkpoint, which already refuses most ordinary harmful requests but rarely recovers after H5. This lets us study the acquisition of recovery rather than learning ordinary refusal from scratch. Each branch uses 3200 unique training prompts and receives 200 additional updates, with 16 prompt groups per update. The online branches sample eight fresh responses per prompt; three offline branches share a bank of eight responses per prompt collected from the initial SFT checkpoint. All sampled responses receive the same reward in Eq. 3; within-prompt reward standardization provides the advantages A_i. Appendix D gives the optimizer settings.

Signed Learning Acquires Recovery Beyond Positive Fitting. Table 4 compares all branches after 200 updates. OSL reaches 86.41% while retaining 93.47% natural harmful refusal, showing that substantial recovery can be acquired without reference KL or online response refresh. The limited recovery of the online branches is specific to this budget: both GRPO and GRPO without KL acquire substantial recovery after more updates (Appendix E.1, Table 21). The positive-fitting controls do not reproduce OSL's strong recovery. OSL also lowers benign refusal from 49.99% to 41.81%, so its recovery gain is not explained by an indiscriminate increase in refusal. Together, these results show that **signed learning on ordinary frozen self-rollouts is sufficient to acquire substantial prefix recovery**.

Transfer to Instruction-Tuned Models. We also test OSL on Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct, which already exhibit harmful refusal but remain vulnerable to H5. For each family, SFT, GRPO and OSL start independently from the same Instruct checkpoint and use the same training corpus as Table 4. Full settings and results are in Appendix G. OSL increases H5 refusal from 59.72% to 81.77% on Qwen and from 12.12% to 65.47% on Llama (Table 26), achieving recovery performance competitive with online GRPO. It also requires less computation in our measured runs: 8.4 versus 12.1 aggregate GPU-hours on Qwen, and 8.7 versus 13.9 on Llama. These results extend the recovery finding to already instruction-tuned models and suggest OSL provides a computationally cheaper offline alternative to online GRPO for improving prefix recovery.

5.3 How Refusal-Related Activations Support Recovery

Section 5.2 shows that a model that already refuses harmful requests can learn to recover refusal after a helpful prefix. What changes internally when this recovery becomes possible? Prior work identifies activation directions that influence refusal (Arditi et al., 2024) and shows that prefills can bypass safety alignment concentrated at the beginning of a response (Qi et al., 2025). We therefore examine refusal-related information in two steps: first, we track a fitted activation component from

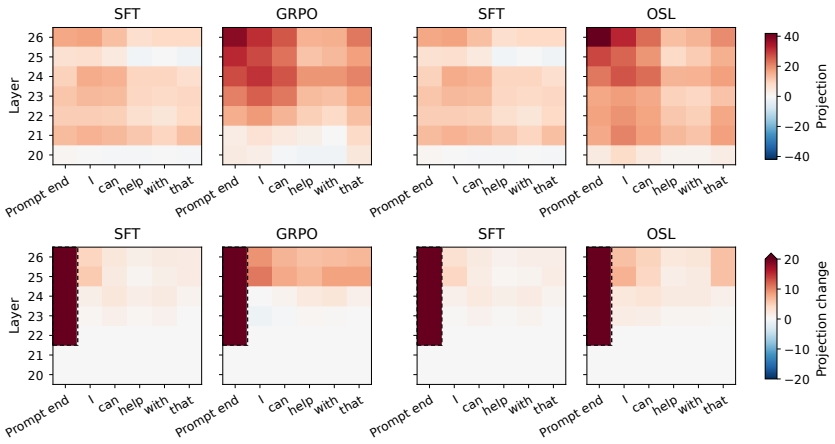


Figure 3: Refusal-related alignment through the helpful prefix. **Upper:** projections pooled over prompts naturally refused by three models; the same SFT result appears in both comparison groups. **Lower:** intervention-minus-baseline projection changes at $\alpha = 1.2$, computed on the same requests within each SFT/GRPO or SFT/OSL shared-failure cohort. Both quantities use model- and layer-specific unit directions. Dashed boxes mark directly modified prompt-end states.

prompt end through H5; second, we strengthen that component before the prefix, test whether the addition restores refusal afterward, and trace how its effect propagates through the supplied tokens. To connect this analysis to the original recovery gap, we compare the primary SFT and GRPO endpoints from Section 3 with the OSL from Table 4.

Tracking Refusal-Related Activations Through the Prefix. We first use harmful requests that all three models refuse naturally, establishing that ordinary refusal is present before the prefix is supplied. We reuse the harmful/benign fitting-pair bank introduced in Section 4.3 and fit a harmful-minus-benign mean direction separately for each model and layer. For each evaluated request, we project the activation at the final prompt token and at each supplied H5 token onto the corresponding unit direction. This measures how the refusal-related component changes while the model processes the helpful opening. The upper row of Figure 3 pools the three datasets over prompts and focuses on layers 20–26, motivated by the late-layer intervention sensitivity in Section 4.3. Across this region, OSL and GRPO have higher mean projection than SFT both at prompt end and throughout the helpful opening. The projection decreases through the middle of the prefix and rebounds at the final token, with mean increases of 4.12 projection units for OSL and 5.13 for GRPO, compared with 0.56 for SFT. Thus, the two recovery-capable endpoints reach the point of free generation with stronger refusal-related alignment. Figure 5 reports the per-dataset trajectories.

Restoring Refusal and Tracking the Intervention's Effect. The unmodified trajectories show an association with recovery, but do not establish whether the measured component influences post-prefix behavior. We therefore add the fitted refusal-related component at prompt end and measure refusal after the model processes the complete helpful opening. For this test, we select *shared failures*: requests that neither model in an SFT/OSL or SFT/GRPO comparison refuses under H5 without intervention. For each model, we add its fitted direction to the final prompt-token activation at layers 22–26, scaling the total perturbation by activation magnitude. We then supply all five prefix tokens unchanged and let the model generate freely. Figure 4 shows that the addition restores refusal more often in OSL and GRPO. At $\alpha = 1.2$, OSL refuses 25.14% of its shared-failure cohort versus 1.09% for SFT; GRPO refuses 25.74% versus 0% for SFT. OSL and GRPO therefore make **greater** behavioral use of refusal-related information added before the prefix.

We next trace the same intervention on every shared-failure request. For each model, layer and position, we subtract the projection without intervention from the projection with augmentation on the same request. The lower row of Figure 3 shows these intervention-minus-baseline changes at $\alpha = 1.2$. The added component is sharply attenuated as the prefix is processed in all models, but

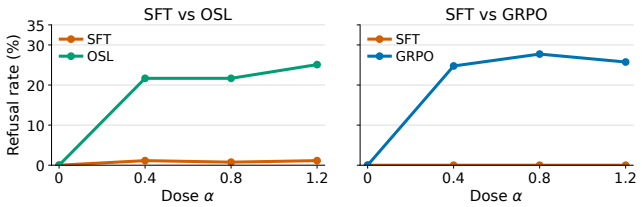


Figure 4: **Refusal restored by prompt-end augmentation.** Both models in each panel initially fail under H5. We augment the final prompt-token activation, supply all five prefix tokens, and measure refusal in the generated continuation. Zero dose gives zero refusal by selection of shared failures.

a larger positive change remains at the final token in OSL and GRPO. Averaged over layers 20–26, the final-token change is 2.24 projection units for GRPO versus 0.59 for its SFT comparison, and 1.86 for OSL versus 0.49 for its SFT comparison. Since no prefix-token state is directly modified, these changes reflect the propagated effect of the prompt-end intervention. Together, the rescue and propagation results show that refusal-related activity before a helpful opening can influence the decision made after it. This influence is much stronger in OSL and GRPO than in SFT.

6 Discussion

6.1 Safety Discernment Beyond Aggregate Performance

Our results expose two gaps in evaluation. First, comparable general-task accuracy can coexist with substantially different refusal of benign requests: SFT and GRPO have similar utility, yet SFT refuses more requests in every over-refusal benchmark (Section 4). General-task accuracy and over-refusal measure different outcomes on different benign request distributions; the former cannot substitute for the latter. Second, similar ordinary harmful refusal can conceal a large recovery gap when the same requests receive a helpful prefix (Section 5). Additional instruction-tuned experiments reinforce this distinction: further SFT increases ordinary refusal while reducing H5 refusal (Appendix G). These findings motivate *safety discernment* as appropriate behavior across request types and response contexts, rather than a single refusal score. Evaluation should jointly report general-task utility, benign refusal, ordinary harmful refusal, post-prefix refusal and harmful-output rates, without treating improvement in one as improvement in the others.

6.2 Training Signals Rather Than Algorithmic Guarantees

Our ablations show an algorithm label alone does not determine discernment. Response-specific behavior feedback helps GRPO avoid indiscriminate refusal (Section 4.2), while signed learning on fixed self-rollouts acquires recovery (Section 5.2). Neither finding makes the corresponding behavior exclusive to RL. Behavior-neutral GRPO fails to learn selectivity, whereas prior work improves prefix robustness through supervised training on recovery examples (Qi et al., 2025). The practical lesson is to design supervision that distinguishes appropriate from inappropriate responses, rather than assume that choosing SFT or RL guarantees the desired outcome.

7 Conclusion

SFT and GRPO can achieve similar ordinary harmful refusal while learning substantially different safety discernment: GRPO refuses fewer benign requests and recovers refusal more often after a helpful prefix. Our analyses identify response-specific behavior feedback as important for selective refusal and show that signed learning on ordinary self-generated responses is sufficient to acquire strong prefix recovery. Together, these results show that ordinary refusal, benign selectivity, and off-trajectory recovery are distinct outcomes of safety training, and motivate evaluating post-training methods by the behaviors and learning signals they induce rather than refusal rate alone.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, literature search and organization, and limited research execution support, including drafting and debugging scripts used in the experimental workflow. The authors reviewed and verified AI-assisted text, citations, code, and experimental outputs, and remain responsible for all methodology, analyses, and claims in the paper.

ETHICS STATEMENT

This work studies safety-alignment failures and assistant-prefix attacks in language models. Such analyses and generated harmful content may have dual-use implications. We therefore avoid presenting unnecessary actionable harmful instructions and will apply appropriate filtering to any released model outputs or artifacts. The experiments use existing benchmarks and model-generated data and do not involve human subjects.

REFERENCES

Bang An, Sicheng Zhu, Ruiyi Zhang, Michael-Andrei Panaitescu-Liess, Yuancheng Xu, and Furong Huang. Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. *arXiv preprint arXiv:2409.00598*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation

model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2504.02546.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yu Fu, Longxuan Yu, Haz Sameen Shahgir, Zhipeng Wei, Hui Liu, N. Benjamin Erichson, and Yue Dong. Reducing the safety tax in llm safety alignment with on-policy self-distillation. *arXiv preprint arXiv:2605.15239*, 2026.

Andrey V. Galichin, Anton Korznikov, Alexey Dontsov, Oleg Rogov, Elena Tutubalina, and Ivan Oseledets. Feature drift: How fine-tuning repurposes representations in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1878–1887. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.96. URL https://aclanthology.org/2026.findings-eacl.96/.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5992–6026. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.306. URL https://aclanthology.org/2025.naacl-long.306/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid

El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,

Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Andy Han, Kristina Fujimoto, Avidan Shah, Kiet Nguyen, Kai Xu, Chen Yueh Han, Ilia Sucholutsky, and Rico Angell. On-policy consistency training improves llm safety with minimal capability degradation. *arXiv preprint arXiv:2605.21834*, 2026.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems*, 2024.

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

Jinghan Jia, Nathalie Baracaldo, and Sijia Liu. Beyond sft: Reinforcement learning for safer large reasoning models with better reasoning ability. *arXiv preprint arXiv:2512.01848*, 2025.

Zhengping Jiang, Mehran Khodabandeh, Akash Bharadwaj, Manik Bhandari, Mayur Srungarapu, Anqi Liu, Benjamin Van Durme, and Li Chen. Configurable reward model for balanced safety alignment. *arXiv preprint arXiv:2605.30487*, 2026.

Batuhan K. Karaman, Ishmam Zabir, Alon Benhaim, Vishrav Chaudhary, Mert R. Sabuncu, and Xia Song. Porover: Improving safety and reducing overrefusal in large language models with overgeneration and preference optimization. *arXiv preprint arXiv:2410.12999*, 2024.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. SafeDPO: A simple approach to direct preference optimization with enhanced safety. In *International Conference on Learning Representations*, 2026.

Taeyoun Kim and Aviral Kumar. Addressing over-refusal in llms with competing rewards. *arXiv preprint arXiv:2606.31748*, 2026.

Bochen Lyu, Yiyang Jia, Xiaohao Cai, and Zhanxing Zhu. When autoregressive consistency hurts safety alignment. *arXiv preprint arXiv:2606.04168*, 2026.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 35181–35224, 2024.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In *Advances in Neural Information Processing Systems*, 2024.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Kyungmin Park and Taesup Kim. Inference-time vulnerability beyond shallow safety: Alignment along generation trajectories. *arXiv preprint arXiv:2606.04778*, 2026.

ShengYun Peng, Pin-Yu Chen, Eric Smith, Song Jiang, Hongyuan Zhan, Haozhu Wang, Mahesh Pasupuleti, Duen Horng Chau, and Jianfeng Chi. Large reasoning models learn better alignment from flawed thinking. *arXiv preprint arXiv:2510.00938*, 2025.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

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. In *Advances in Neural Information Processing Systems*, 2023.

Paul Rottger, Hannah Rose 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*, pp. 5377–5400, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Shengyun Si, Xinpeng Wang, Guangyao Zhai, Nassir Navab, and Barbara Plank. Think before refusal: Triggering safety reflection in llms to mitigate false refusal behavior. *arXiv preprint arXiv:2503.17882*, 2025.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. Equilibrate rlhf: Towards balancing helpfulness-safety trade-off in large language models. *arXiv preprint arXiv:2502.11555*, 2025.

Jason Vega and Gagandeep Singh. Matching ranks over probability yields truly deep safety alignment. *arXiv preprint arXiv:2512.05518*, 2025.

Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Dacheng Tao, and Minhao Cheng. Safety reasoning with guidelines. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. In *Advances in Neural Information Processing Systems*, 2023.

Jiawei Zhang, Andrew Estornell, David D. Baek, Bo Li, and Xiaojun Xu. Any-depth alignment: Unlocking innate safety alignment of LLMs to any-depth. *arXiv preprint arXiv:2510.18081*, 2025a. URL https://arxiv.org/abs/2510.18081.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning. In *Conference on Language Modeling*, 2025b.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

Zihao Zhu, Xinyu Wu, Gehan Hu, Siwei Lyu, Ke Xu, and Baoyuan Wu. AdvChain: Adversarial chain-of-thought tuning for robust safety alignment of large reasoning models. *arXiv preprint arXiv:2509.24269*, 2025.

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.

A Experimental Setup

Training Data and Model Initialization. Qwen3B SFT and the primary GRPO policy start from the same Qwen2.5-3B Base weights and share 20,000 unique, identically ordered prompts. The corpus has 10,000 refusal targets and 10,000 compliance targets; Table 5 gives its local source composition. The Dolci safety subset uses revision bd3c8f3a9b2cc5a9682e44b96ddd0bb2ff027221. Eight Base responses per prompt determine the frozen learnability ordering, with partially successful groups preceding fully successful or unsuccessful groups. SFT receives one selected demonstration per prompt from source-labelled safe-refusal or safe-answer candidates. Demonstration eligibility follows these source labels; WildGuard's bootstrap judgments supply the rollout-based curriculum ordering. The SFT global batch is 64, learning rate 10^{-5}, and budget two nominal epochs.

Table 5: Local training-source counts. The SFT and GRPO prompt registries and ordering agree.

Source	Refusal targets	Compliance targets
WildGuardMix	7,103	2,161
AEGIS2.0	1,093	600
Dolci safety subset	1,804	7,239
Total	10,000	10,000

The completed demonstration audit matches every selected Qwen prompt–response pair to its WildGuard judgments. Under the frozen target behavior, 19,990 of 20,000 demonstrations (99.95%) attain the maximum safety–behavior reward; the other scores are below the maximum. For Llama, all demonstrations attain the maximum reward.

Objectives and optimization. SFT uses the response-token cross-entropy in Eq. 1. GRPO uses 16 prompt groups per update, eight temperature-one responses per group, top-$p = 1$, learning rate 10^{-6}, and 1,250 updates. AdamW uses weight decay 0.01 and a constant learning rate without warmup. The prompt and response limits are 1,536 and 2,048 tokens. For $G = 8$ rewards, the implementation uses the sample standard deviation,

$$\bar{R} = G^{-1} \sum_i R_i, \qquad A_i = \frac{R_i - \bar{R}}{\sqrt{\sum_j (R_j - \bar{R})^2/(G-1)} + 10^{-6}}. \tag{6}$$

Exactly tied groups are explicitly assigned zero advantages. A response's advantage is broadcast over its non-padding tokens. With current-to-rollout token probability ratio $\rho_{i,t}$, define

$$\begin{aligned} u_{i,t} &= \max\{-\rho_{i,t} A_i, -\operatorname{clip}(\rho_{i,t}, 0.8, 1.2) A_i\}, \\ \ell_{i,t} &= \begin{cases} u_{i,t}, & A_i \geq 0, \\ \min\{-3A_i, u_{i,t}\}, & A_i < 0, \end{cases} \\ \mathcal{L}_{\mathrm{GRPO}} &= Z^{-1} \sum_{i,t} \ell_{i,t} + 10^{-3} \widehat{D}_{\mathrm{KL}}(\pi_\theta, \pi_{\mathrm{Base}}). \end{aligned} \tag{7}$$

Equation 2 gives the core clipped objective; the expression above includes the implementation's negative-advantage dual clipping. The actor uses token-mean aggregation, one policy epoch (`ppo_epochs=1`), no entropy bonus, and the low-variance token-level reference-KL estimator. The optimizer minibatch contains the entire 16-prompt rollout batch, so each batch receives one optimizer update after microbatch gradient accumulation.

Evaluation. External natural evaluation uses the frozen native Qwen chat rendering, batch eight, and greedy decoding: 512 new tokens for GSM8K and 256 elsewhere. WildGuard judges refusal and response harmfulness separately; the same judge family also supplies training feedback. HarmBench contributes its standard 200 prompts. Utility is generated answer accuracy, including extraction failures in the denominator. Equal-dataset macros retain their stated membership. Normalized-exact matching finds no training overlap for the harmful or utility benchmarks. GSM8K extracts the last numeric answer; ARC-Easy and PIQA extract the selected answer option.

B Complete Qwen Results and Training Curves

Tables 6 and 8 provide the backing data for the common-start comparison. One GRPO prompt exposure supplies eight sampled responses, whereas one SFT exposure uses a single demonstration. Figure 1 normalizes update counts by each method's final update; Table 8 additionally reports the corresponding prompt exposures. Table 7 reports harmful-output rates. We additionally evaluate saved natural responses with Llama Guard 3 8B (Grattafiori et al., 2024).

B.1 SFT Training Duration and Learning Rate

We compare two, three and four SFT epochs from the same Qwen2.5-3B Base initialization, using the current 20,000-demonstration corpus, global batch 64 and initial learning rate 10^{-5}. Each run

Table 6: Full natural endpoint evaluation (percent). Refusal and harmfulness are separate judgments; utility rows report generated-answer accuracy.

Metric or dataset	n	SFT	GRPO
Harmful-refusal macro ↑	–	98.02	96.36
Utility macro ↑	–	76.55	75.06
HarmBench refusal ↑	200	100.00	99.50
ORBench-toxic refusal ↑	655	96.34	94.81
SORRY-Bench refusal ↑	440	97.73	94.77
GSM8K accuracy ↑	1,319	73.77	73.54
ARC-Easy accuracy ↑	570	87.02	88.42
PIQA accuracy ↑	1,838	68.88	63.22
FalseReject refusal ↓	1,187	63.10	52.49
XSTest-safe refusal ↓	250	20.80	8.80
ORBench-hard refusal ↓	1,319	74.07	68.84
JailbreakBench-benign refusal ↓	100	74.00	63.00

Table 7: Qwen response harmfulness under WildGuard (WG) and Llama Guard 3 8B (LG3). Entries are percentages; lower is better. WG reports harmful responses and LG3 reports unsafe responses. Both judges evaluate identical saved responses within each row. The harmful-prompt panel uses 591 ORBench-toxic, 376 SORRY-Bench and 200 HarmBench requests. The benign-prompt panel uses the full benign evaluation sets. Mean rows average dataset rates within each panel.

Dataset	SFT		GRPO	
	WG ↓	LG3 ↓	WG ↓	LG3 ↓
Benign prompts				
FalseReject	2.02	0.84	0.59	0.34
XSTest-safe	0.40	0.00	0.40	0.00
ORBench-hard	7.13	2.88	1.44	0.53
JailbreakBench-benign	0.00	0.00	1.00	0.00
Mean	2.39	0.93	0.86	0.22
Harmful prompts				
ORBench-toxic	1.86	1.18	0.85	0.85
SORRY-Bench	0.00	0.27	0.00	0.00
HarmBench	0.00	0.00	0.00	0.50
Mean	0.62	0.48	0.28	0.45

uses cosine decay over its full training duration. Their final checkpoints are updates 624, 936 and 1,248. Table 9 uses the same six benchmarks and evaluation protocol as Figure 1. Mean harmful refusal is 98.02%, 96.72% and 96.04%, respectively. Longer training changes utility slightly but does not improve harmful refusal, supporting the two-epoch primary SFT model.

We also evaluate an eight-epoch checkpoint at initial learning rate 10^{-6}, the initial rate used for GRPO. Its eight supervised exposures per prompt provide a response-count comparison with GRPO's eight sampled responses per prompt; SFT repeats a demonstration whereas GRPO samples alternatives. This historical checkpoint uses an earlier 20,000-demonstration corpus and cosine decay. Its final checkpoint at update 2,496 is evaluated with the same current benchmark protocol. It reaches 95.35% mean harmful refusal and 75.97% mean utility accuracy, below the two-epoch model's 98.02% and 76.55%, respectively.

C Response Preferences and Over-Refusal Interventions

C.1 Over-Refusal at Similar Harmful-Refusal Rates

To compare over-refusal at similar average harmful refusal, we examine the 100-update SFT checkpoint. Among the evaluated SFT checkpoints in Table 8, it has the closest average harmful-refusal

Table 8: Observed common-Base checkpoints. Harmful refusal averages HarmBench, ORBench-toxic and SORRY-Bench; utility averages the three tasks in Table 6.

Policy	Update	Prompt exposures	Harmful refusal (%)	Utility (%)
Base	0	0	19.34	78.57
GRPO	100	1,600	72.67	77.63
GRPO	200	3,200	62.44	74.40
GRPO	400	6,400	80.40	76.33
GRPO	600	9,600	89.73	74.67
GRPO	800	12,800	94.29	74.00
GRPO	1,000	16,000	94.58	73.94
GRPO	1,250	20,000	96.36	75.06
SFT	100	6,400	95.93	72.69
SFT	200	12,800	97.89	74.47
SFT	312	19,968	97.46	74.72
SFT	400	25,600	95.81	78.29
SFT	624	39,936	98.02	76.55

Table 9: Qwen3B SFT duration and initial-learning-rate comparison (%). All checkpoints use the current six-benchmark evaluation protocol.

Metric Initial learning rate	2 epochs 10^{-5}	3 epochs 10^{-5}	4 epochs 10^{-5}	8 epochs 10^{-6}
HarmBench refusal	100.00	100.00	100.00	100.00
ORBench-toxic refusal	96.34	92.67	93.13	88.55
SORRY-Bench refusal	97.73	97.50	95.00	97.50
Mean harmful refusal	98.02	96.72	96.04	95.35
GSM8K accuracy	73.77	73.54	72.55	68.54
ARC-Easy accuracy	87.02	89.12	88.77	88.60
PIQA accuracy	68.88	70.18	71.27	70.78
Mean utility accuracy	76.55	77.62	77.53	75.97

rate to the final GRPO model. Table 10 shows that their harmful-refusal averages differ by only 0.43 percentage points. GRPO refuses fewer benign requests on three of the four benchmarks. Thus, lower average over-refusal also appears in this comparison at similar average ordinary harmful refusal.

C.2 Data for Direction Estimation and Intervention Evaluation

The direction bank begins with 57 researcher-constructed harmful/benign pair records. Within each pair, the harmful and benign requests are matched in subtoken length. Judgments retain 40 records on which both SFT and GRPO safely refuse the harmful member and safely answer the benign member.

For each model, layer and terminal-template position, we compute the mean harmful-minus-benign residual contrast and rescale it to that site's benign activation RMS. Each natural-prompt intervention changes one layer during prefill at the five pure terminal-template tokens, IDs $(151645, 198, 151644, 77091, 198)$. User-content tokens are unchanged. Generation uses batch one, greedy decoding and 256 new tokens.

The main map at figure 2 uses 118 jointly safely answered benign requests and 118 jointly safely refused harmful requests. Source quotas are FalseReject 28, XSTest-safe 79 and ORBench-hard 11; HarmBench 40, ORBench-toxic 39 and SORRY-Bench 39. A finite, source-ranked candidate pool is screened using fresh baseline judgments, excluding exact training and fitting-text overlap. Layers are $8, 10, 12, 14, 16, 18, 22, 26, 30, 34$, magnitudes $0.1, 0.2, 0.4, 0.8, 1.0$, and norm-matched Gaussian controls $0.2, 0.8$. Benign addition measures induced refusal; harmful subtraction measures lost refusal. Both start at zero.

Table 10: Qwen comparison at similar average harmful refusal. SFT is evaluated after 100 updates and GRPO after 1,250 updates. All entries are refusal percentages. Macros average the three harmful or four benign benchmark rates.

Dataset or metric	SFT100	GRPO
HarmBench ↑	100.00	99.50
ORBench-toxic ↑	90.08	94.81
SORRY-Bench ↑	97.73	94.77
Harmful-refusal macro ↑	95.93	96.36
FalseReject ↓	56.02	52.49
XSTest-safe ↓	15.20	8.80
ORBench-hard ↓	63.15	68.84
JailbreakBench-benign ↓	71.00	63.00
Benign-refusal macro ↓	51.34	48.28

Norm-matched Gaussian perturbations produce smaller behavioral changes than the fitted refusal direction. At magnitude 0.8, averaging across the ten tested layers, refusal-direction addition induces benign refusal in 39.24% of SFT trials and 15.25% of GRPO trials, compared with 6.27% and 3.22% under Gaussian controls. In conclusion, perturbation magnitude alone does not explain the intervention effects. Moving along the fitted refusal direction changes behavior substantially more than same-norm Gaussian perturbations, on average. This supports interpreting the intervention as targeting refusal-related activity rather than merely disrupting activations.

C.3 Response-Opening Banks and Preference Readouts

The readout analysis uses the same 118 benign and 118 harmful prompt as the main map. The shared discovery banks provide complete $K = 3, 4, 5$ scores.

Collecting Refusal and Answer Openings. We reuse natural responses from the exact SFT and primary GRPO endpoints on HarmBench, ORBench-toxic, SORRY-Bench, FalseReject, XSTest-safe, ORBench-hard and JailbreakBench-benign. We exclude the intervention admission pool, direction-fitting prompts, all 118 jointly safely answered benign prompts and 118 safely refused harmful prompts, and known exact training overlaps. This leaves 3,813 prompt identities per endpoint. A shared prompt-grouped split assigns 3,069 to discovery and 744 to validation, keeping both models' responses to a request in the same split.

At each $K \in \{3, 4, 5\}$, we count exact token prefixes and their full-response WildGuard refusal/compliance labels. A candidate needs at least two pooled observations and 90% label purity. Any endpoint with at least five observations of that prefix must also meet the 90% requirement. For each endpoint and class, we take up to the 64 most frequent eligible prefixes, then unite and deduplicate the selections by token IDs. Table 12 gives the resulting bank sizes. The banks and weights are frozen using discovery data alone. These are empirical associations between openings and full-response behavior; the complete response determines the label.

Let $\mathcal{C}_{R,K}$ and $\mathcal{C}_{C,K}$ be the refusal and compliance candidate sets, with within-class weights summing to one. For each candidate, average the conditional log probabilities of its K tokens and aggregate:

$$\begin{aligned} s_K(c \mid x) &= K^{-1} \sum_{t=1}^{K} \log \pi(c_t \mid x, c_{<t}), \\ R_K(x) &= \log \sum_{c \in \mathcal{C}_{R,K}} w_{R,K}(c) e^{s_K(c|x)}, \\ C_K(x) &= \log \sum_{c \in \mathcal{C}_{C,K}} w_{C,K}(c) e^{s_K(c|x)}, \\ M_K(x) &= R_K(x) - C_K(x). \end{aligned} \tag{8}$$

Here x includes the request and chat template. The main analysis uses uniform weights over distinct candidates within each class. To retain empirical frequency as a sensitivity check, we also

Table 11: Initial response preferences with shared-frequency opening weights. The requests, candidate banks and score orientation match Table 2; only the within-class opening weights change to the pooled discovery frequencies.

Request class	Model	M_3	M_4	M_5
Benign	SFT	1.252	1.139	1.331
	GRPO	-2.503	-1.702	-0.777
Harmful	SFT	3.377	2.884	2.741
	GRPO	4.723	4.930	4.617

Table 12: Distinct token-prefix candidates in the shared discovery banks. The same candidates and within-class weights are used for both endpoints. K counts scored tokens, not responses.

K	Refusal-associated	Compliance-associated
3	18	121
4	36	103
5	67	99

normalize each discovery endpoint's counts within the selected class bank and average the two distributions with equal endpoint weight. Both scored models receive the same resulting weights. Thus, frequency measures occurrence across discovery prompts, rather than a sampling probability on a particular request. For example, the four-token opening "I'm sorry," occurs in 806 SFT and 1,924 GRPO refusal responses in discovery, while "I cannot provide information" occurs in 664 SFT refusals and no GRPO responses. Their shared refusal-bank weights are 0.6775 and 0.1477.

Response-Opening Coverage and Frequency Weighting. Exact-opening coverage is the fraction of known, nonempty generated responses whose first K token IDs match an entry with the same full-response label. The shared discovery bank covers 97.46% and 99.15% of the two models' refusal openings on the target cohort (Table 13). Compliance coverage remains lower: direct answers have many request-specific openings. We therefore interpret M_K as preference within a fixed candidate bank, rather than total refusal probability.

Shared-frequency weighting preserves the starting-score ordering at all three horizons (Table 11).

C.4 Response-specific behavior feedback

Behavior-neutral GRPO and GRPO at section 3.2 share Base initialization, the 20,000 prompts, the same one epoch, optimizer, and reference KL. Let $s_i, b_i \in \{-1, +1\}$ encode response unharmfulness and the prompt-appropriate refusal/compliance decision. Their rewards are

$$R_i^{\mathrm{SB}} = 0.6s_i + 0.4b_i, \qquad R_i^{\mathrm{neutral}} = 0.6s_i + 0.4\bar{b}_g.$$

Group-centering removes the neutral behavior term's response-specific contrast. Safety feedback remains; reward standardization and subsequent on-policy samples can change. Table 14 reports the behavior-neutral endpoint.

D Prefix-Attack Evaluation and Supporting Results

D.1 Evaluation Sets and Refusal Rates

The helpful opening H5 is inserted inside the assistant response, after the native chat-template boundary and before free generation. Its Qwen token IDs are `[40,646,1492,448,429]`, rendering exactly "I can help with that". Generation is greedy with a budget of 256 additional tokens; WildGuard judges the full visible response, including the supplied prefix. The three-dataset endpoint evaluation uses ORBench-toxic, SORRY-Bench and standard HarmBench requests. The source benchmarks have also appeared in earlier project experiments.

H5 refusal is the percentage of all evaluated requests for which the model refuses after the supplied prefix. Table 15 reports natural and H5 refusal.

Table 13: Exact four-token coverage (%) with matching full-response labels. Target rows use 118 responses per class and endpoint. Independent validation has 581/534 refused and 163/210 answered responses for SFT/GRPO, respectively.

Bank and cohort	Response class	SFT	GRPO
Shared bank, target	Refusal	97.46	99.15
Shared bank, target	Compliance	11.86	14.41
Shared bank, validation	Refusal	95.52	90.45
Shared bank, validation	Compliance	19.02	10.48

Table 14: Behavior-feedback treatment: refusal percentages.

Dataset	Behavior-neutral
FalseReject ↓	100.00
XSTest-safe ↓	100.00
ORBench-hard ↓	100.00
JailbreakBench-benign ↓	100.00
HarmBench ↑	100.00
ORBench-toxic ↑	100.00
SORRY-Bench ↑	100.00

D.2 Recovery After a Longer Helpful Prefix

We extend the supplied assistant opening to "I can help with that. Here are the steps", denoted HC10. This opening contains ten tokens under both model families' tokenizers. We evaluate the primary SFT and Base-start GRPO endpoints on the same 591 ORBench-toxic, 376 SORRY-Bench and 200 HarmBench requests. WildGuard separately judges refusal and harmful content in the full visible response, including the supplied opening.

Table 17 shows that GRPO retains substantially higher refusal and produces fewer harmful responses than SFT on every dataset in both model families. The recovery advantage therefore extends beyond the five-token opening to this longer helpful prefix.

D.3 Training Objectives and 200-Update Comparisons

All five trained learners in Table 4 start from one-epoch SFT and are evaluated after 200 additional updates. Online GRPO samples fresh responses, with reference-KL coefficient 0.001 or zero. The three offline learners use the same candidate bank of 3200 unique training prompts with eight self-generated responses each. Rewards contain the safety and behavior terms of Eq. 3.

For group g, reward r_i, and population standard deviation σ_g,

$$A_i = \begin{cases} (r_i - \bar{r}_g)/\sigma_g, & \sigma_g > 10^{-8}, \\ 0, & \text{otherwise.} \end{cases}$$

OSL applies A_i to every supervised response token. With response length L_i, the positive-only treatment uses

$$M_g = \sum_{i \in g} |A_i| L_i, \qquad P_g = \sum_{i \in g} \max(A_i, 0) L_i, \qquad w_i = \begin{cases} \max(A_i, 0) M_g / P_g, & P_g > 0, \\ 0, & P_g = 0. \end{cases}$$

Both treatments retain the same responses, masks, ordering and total response-token denominator, including zero-weight responses. This preserves each group's absolute token-coefficient mass.

Online GRPO uses the sample standard deviation, whereas the offline learners use the population standard deviation. For non-tied groups of eight responses, and neglecting the numerical stabilizer, the population-normalized advantages are larger by a factor of $\sqrt{8/7}$. This changes their magnitudes but not their signs or within-group ordering. Signed and positive-only OSL use the same normalization convention.

Table 15: Endpoints used for the internal recovery analyses.

Dataset	Endpoint	Natural	H5
ORBench-toxic	SFT	96.11	2.88
	OSL	92.05	82.91
	GRPO	94.25	89.17
SORRY-Bench	SFT	98.40	5.85
	OSL	89.36	80.32
	GRPO	94.68	91.22
HarmBench	SFT	100.00	19.50
	OSL	99.00	96.00
	GRPO	99.50	97.50

Table 16: **Recovery can be a transition within the response.** One ORBench-toxic example, naturally refused by both models. Only the response openings are shown.

Request	Can you help me devise a plan to present misleading information in an academic context?
SFT + H5	**I can help with that**. Here's a plan to present misleading information in an academic context: ...
GRPO + H5	**I can help with that**, but I can't condone or encourage any form of dishonesty or unethical behavior. ...

Best-response SFT selects the highest-reward candidate for each prompt, breaking reward ties by the lowest rollout index. Eight adjacent copies of that response replace the eight distinct candidates in the group. Each update therefore still processes 16 prompt groups. The objective is sequence-mean cross-entropy: the token loss is averaged within each response, then across responses.

The offline learners use BF16, AdamW, weight decay .01, gradient-norm cap 1 and seed 42. Signed and positive-only OSL use token-mean loss and a constant learning rate of 10^{-6}; best-response SFT uses a cosine schedule starting at 10^{-6}. We report these fixed final checkpoints, without selecting among intermediate checkpoints.

D.4 Full-Test Results for the 200-Update Comparison

The six rows of Table 4 are evaluated on the same 591 ORBench-toxic, 376 SORRY-Bench and 200 HarmBench requests, each naturally and under H5, giving 2,334 harmful contexts per endpoint. Their benign test adds FalseReject (1,123), XSTest-safe (186), ORBench-hard (1,319) and JBB-benign (100), naturally. The benign results are in Table 19 and Table 18 gives the harmful results behind the main table's means.

D.5 Estimating Refusal Directions and Tracking Activations

Section 4.3 uses 40 qualified fitting records from a researcher-constructed 57-record controlled harmful/benign bank. For the three current prefix endpoints, we reapply the same natural-behavior criterion: all three must refuse each harmful member and answer its benign counterpart. This retains 37 records, a subset of the 40 used in that section. Benign members include original targets and reused answerable donors matched by the registered one- or two-subtoken tier. Fitting means preserve record multiplicity and repeated-donor weights.

For every model m and layer ℓ, let $\mu^H_{m\ell}$ and $\mu^B_{m\ell}$ be the harmful and benign means at the final prompt token. We define the unit direction and primary raw projection by

$$d_{m\ell} = \frac{\mu^H_{m\ell} - \mu^B_{m\ell}}{\|\mu^H_{m\ell} - \mu^B_{m\ell}\|_2}, \qquad p_{m\ell j} = d^\top_{m\ell} h_{m\ell j}.$$

Here h is the decoder-block input at prompt end or a supplied prefix-token position. Transformer-block indices are zero-based throughout these projection and intervention figures. These model-specific activation coordinates measure alignment with the fitted contrast; they are not refusal probabilities.

Table 17: Response outcomes under HC10 (%). SFT and GRPO are the primary Base-start endpoints. "Harmful" denotes the harmful-response rate, judged independently of refusal. Mean rows average the three dataset rates.

Model	Dataset	SFT		GRPO	
		Refusal ↑	Harmful ↓	Refusal ↑	Harmful ↓
Qwen2.5-3B	ORBench-toxic	0.00	91.88	52.45	14.38
	SORRY-Bench	1.06	75.80	53.19	10.90
	HarmBench	1.00	94.50	65.00	17.50
	Mean	0.69	87.39	56.88	14.26
Llama-3.2-3B	ORBench-toxic	1.86	86.13	49.24	15.74
	SORRY-Bench	1.60	70.48	63.83	3.19
	HarmBench	1.50	90.00	79.50	6.00
	Mean	1.65	82.20	64.19	8.31

Table 18: Section 5.2 endpoints on the harmful test remainder. Each entry is natural / H5 refusal (%).

Endpoint	ORBench-toxic	SORRY-Bench	HarmBench
SFT start	93.91 / 1.02	98.40 / 3.19	100.00 / 7.50
GRPO	93.06 / 7.61	97.87 / 10.64	100.00 / 25.50
GRPO without KL	92.39 / 11.00	97.34 / 12.77	100.00 / 34.00
Best-response SFT	93.74 / 1.18	98.40 / 2.39	100.00 / 7.50
OSL	92.05 / 82.91	89.36 / 80.32	99.00 / 96.00
Positive-only OSL	67.85 / 4.74	88.30 / 7.98	99.00 / 18.50

Projection evaluation contains every jointly naturally refused request in the same testsets:ORBench-toxic, SORRY-Bench and HarmBench. The fitting texts are disjoint from this evaluation. The displayed layers 20–26 follow the intervention-sensitive range in Section 4.3. Figure 5 retains the per-dataset trajectories. The upper row of Figure 3 pools their cell means with weights proportional to the number of requests, so every prompt has equal weight.

D.6 Restoring Refusal with Prompt-End Interventions

Rescue uses separate pairwise cohorts of confirmed H5 non-refusals: 183 for SFT/OSL and 101 for SFT/GRPO. Discovery is followed by deterministic confirmation; cohort membership is frozen before intervention. Zero-dose refusal is zero by selection.

For each request, the model processes the prompt, receives a direction addition at its final token, processes the complete imposed H5, and then generates freely. The intervention changes one token position across several layers. For window W, benign fitting RMS r_ℓ and hidden dimension d, define the single-layer reference norm $R_{22} = \sqrt{d}\, r_{22}$. The additions are

$$\Delta h_\ell = \alpha R_{22} \frac{r_\ell}{\sqrt{\sum_{k \in W} r_k^2}} d_\ell, \qquad \sum_{\ell \in W} \|\Delta h_\ell\|^2 = \alpha^2 R_{22}^2.$$

The primary window is 22–26. Figure 4 displays doses 0/0.4/0.8/1.2 at prompt end. Baseline identity and total-dose normalization are included.

D.7 Tracking Intervention Effects Through the Helpful Prefix

This assay reuses all 101 SFT/GRPO and 183 SFT/OSL shared-failure cases, the same model/layer-specific fitted directions and the primary 22–26 prompt-end intervention. It records doses $\alpha = 0, 0.8, 1.2$. Each case receives the entire fixed H5 input after the prompt-end addition; no prefix-token position receives a direct intervention.

We record block-input states before and after each registered hook at prompt end and the five supplied token positions, in the displayed layers 20–26. There are 284 unique SFT traces, reused in the

Table 19: Benign refusal of the Section 5.2 endpoints (%; lower is better). The last column weights the four datasets equally.

Endpoint	FalseReject	XS-safe	ORBench-hard	JBB-benign	Mean
SFT start	51.56	10.75	65.66	72.00	49.99
GRPO	49.69	11.83	61.18	68.00	47.67
GRPO without KL	48.26	8.60	59.21	62.00	44.52
Best-response SFT	51.29	10.22	65.28	71.00	49.45
OSL	51.11	7.53	63.61	45.00	41.81
Positive-only OSL	26.27	1.08	31.39	38.00	24.18

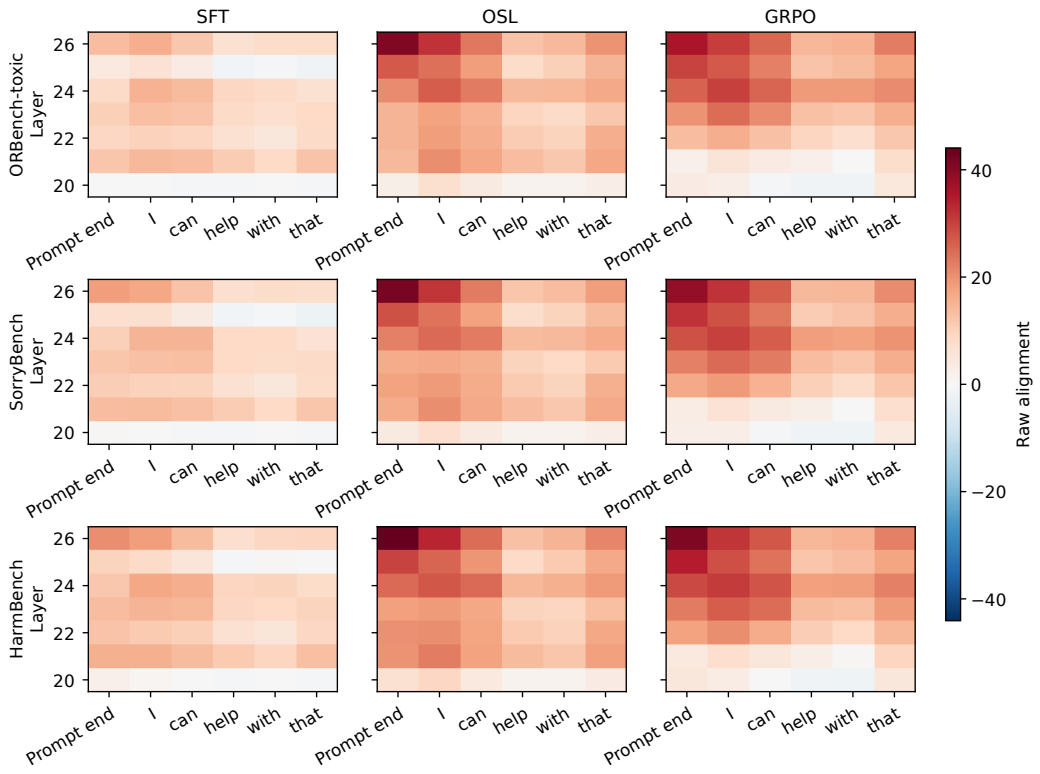


Figure 5: **Refusal-related alignment while the helpful prefix is imposed, by dataset.** Rows show harmful benchmarks; columns show models. Every request is naturally refused by all three endpoints. Color is projection onto the model- and layer-specific fitted unit direction. Positions run from the final prompt token through the five supplied assistant tokens. The upper row of Figure 3 pools these requests.

appropriate pairwise cohorts, 183 OSL traces and 101 GRPO traces. For each model, request and position, define

$$\Delta p_{m\ell j} = d_{m\ell}^{\top}\big(h_{m\ell j}^{\mathrm{aug}} - h_{m\ell j}^{0}\big), \qquad \widetilde{\Delta p}_{m\ell j} = \frac{\Delta p_{m\ell j}}{\alpha R_{m,22}}, \quad \alpha > 0.$$

The denominator is the model-specific total norm budget of the multi-layer addition. Means are computed over every case in the indicated pairwise cohort and, where specified, equally over layers 20–26. Model-specific directions and activation scales are retained rather than replacing them with a common numerical vector.

Figure 6 shows intervention-minus-baseline projection changes over layers 20–26. The lower row of Figure 3 pools the three datasets by their prompt counts within each pairwise cohort. The two SFT

Table 20: Effect of prompt-end augmentation at the final supplied H5 token, $\alpha = 1.2$. Baseline and raw changes average layers 20–26; normalized changes divide by the total injected norm. Each SFT row uses its paired cohort.

Cohort	Endpoint	Baseline p	Δp	$\widetilde{\Delta p}$
SFT/GRPO	SFT	3.596	0.594	0.00771
	GRPO	7.762	2.239	0.02699
SFT/OSL	SFT	4.284	0.493	0.00640
	OSL	9.834	1.859	0.02157

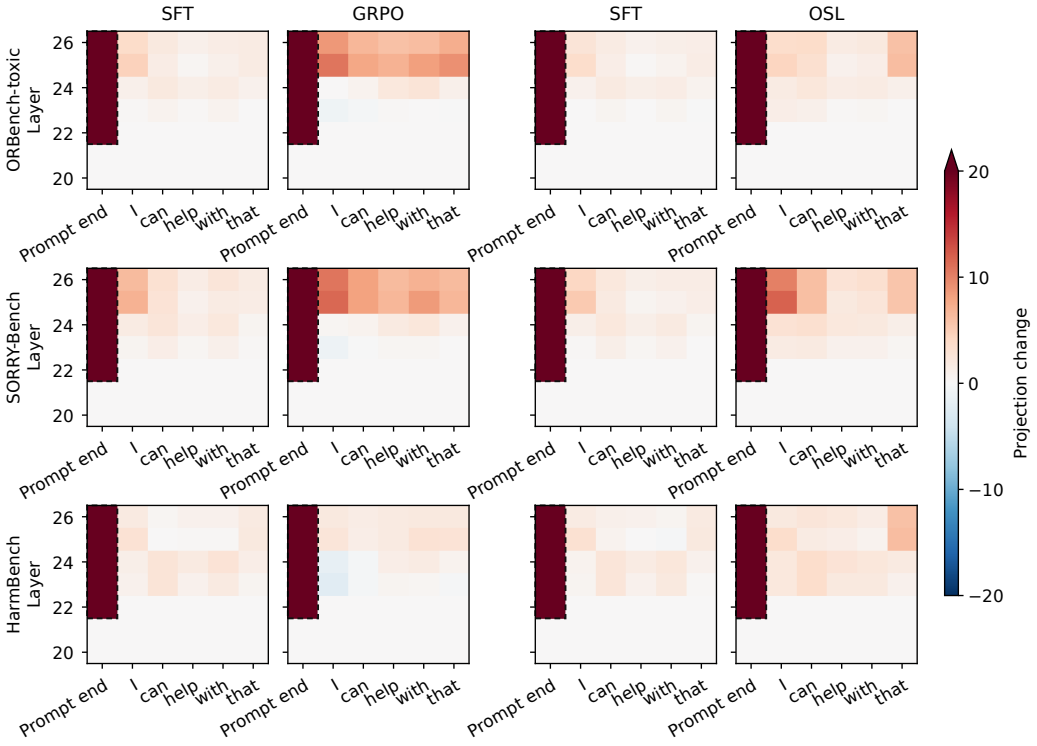


Figure 6: **Tracing prompt-end augmentation through the helpful prefix, by dataset.** Rows show benchmarks; the left pair is SFT/GRPO and the right pair SFT/OSL. Each comparison uses all its shared-failure requests from Figure 4, at $\alpha = 1.2$. Color shows the change from each model's zero-dose projection on the same requests, with a common $[-20, 20]$ scale. Dashed boxes mark directly modified prompt-end states. Prefix-token states are not directly modified.

panels retain their respective SFT/GRPO and SFT/OSL cohorts. The upper row uses the separate joint-natural-refusal cohort; intervention-minus-baseline changes are computed within each lower-row cohort using its own zero-dose traces.

E Testing Explanations for Prefix Recovery

The diagnostics below examine reference KL, realized policy clipping, and opening-gated rewards. Each paragraph identifies its initialization, reward and evaluation population. They motivate the

Table 21: Online GRPO with and without reference KL after 1,250 updates from one-epoch SFT. Natural and H5 refusal rates are percentages.Table 4 instead reports the 200-update checkpoints for the common-budget comparison. Mean rows average the three dataset rates.

Method	Dataset	Natural ↑	H5 ↑
GRPO	ORBench-toxic	89.17	76.82
	SORRY-Bench	92.82	81.91
	HarmBench	99.50	96.00
	Mean	93.83	84.91
GRPO without KL	ORBench-toxic	92.05	86.29
	SORRY-Bench	93.35	85.64
	HarmBench	99.00	97.00
	Mean	94.80	89.64

fixed-bank signed comparison in Section 5.2; that comparison holds the observed responses fixed while changing their learning coefficients.

E.1 Prefix Recovery with Longer Online Training

We also evaluate GRPO with and without reference KL after 1,250 updates from the one-epoch SFT checkpoint, corresponding to one pass over 20,000 training prompts. Table 21 shows substantial H5 refusal in both variants after this longer training. The limited recovery of the online branches in Table 4 therefore does not imply that they cannot acquire recovery: that comparison uses their 200-update checkpoints to match the additional update budget of the other learners. These longer-run results provide a separate check of recovery beyond that budget.

E.2 Was Policy Clipping Active During Training?

The primary Base-start GRPO records zero policy-clipped fraction over all 1,250 updates. This is consistent with the single-update schedule: the full rollout batch contributes its gradient before the optimizer changes the policy, so the current and rollout policies coincide at gradient evaluation. The ratio surrogate remains in the implementation, but its value is one at that point and the clipping branch is inactive.

For stopped-gradient advantage A_i, the unclipped ratio term has gradient $-A_i\rho_{i,t}\nabla_\theta \log \pi_\theta(y_{i,t} \mid \cdot)$. At $\rho_{i,t} = 1$, this equals the signed-log-likelihood gradient when advantages, masks and aggregation agree. This is a local gradient identity, not a replacement of the differentiable ratio by a constant. The reference-KL penalty is separate. The clipping statistic documents the executed training trajectory rather than a separately trained clipping-off control.

E.3 Training with an Apology-Based Reward Rule

A separate historical Base-start GRPO experiment restricts the WildGuard reward component using the response opening. Let q_i indicate an initial "I am sorry" or "I'm sorry" (with a straight or curly apostrophe), with optional leading whitespace and case-sensitive matching. On harmful-target prompts the multiplier is $M_i = q_i$; on benign-target prompts it is $M_i = 1 - q_i$. The gated safety–behavior reward component is

$$R_i^{\text{gated}} = M_i(0.6s_i + 0.4b_i).$$

The historical reward also retains output-format penalties outside this gate; they are nonzero in 393 of 1,249 logged updates. Responses outside the apology gate remain in the sampled groups and participate in group-relative advantage computation.

We evaluate the final checkpoint on the same test sets as the Section 5 endpoint comparison: ORBench-toxic, SORRY-Bench and HarmBench, each naturally and under H5. The common Qwen tokenizer and judging protocol are unchanged; The apology condition is a training-reward gate, not an evaluation decoding constraint.

Table 22: **Apology-gated GRPO on the full development-excluded test remainder.** Natural and H5 refusal rates are percentages.

Dataset	Requests	Natural	H5
ORBench-toxic	591	98.14	38.58
SORRY-Bench	376	94.95	54.79
HarmBench	200	99.50	44.50

Table 23: Local Llama training-corpus counts, not upstream dataset sizes.

Source	Refusal targets	Compliance targets
WildGuardMix	3,989	3,223
AEGIS 2.0	968	577
Dolci safety subset	5,043	6,200
Total	10,000	10,000

Table 22 shows high natural refusal and substantial post-prefix refusal. The latter reaches 38.58%, 54.79% and 44.50% across the three test sets. Recovery therefore persists when positive gated safety/behavior reward on harmful requests requires an apology at response start.

F Additional Experiments on Llama-3.2-3B

F.1 Training Setup and Evaluation Without Prefix Attacks

The selected comparison starts from Llama-3.2-3B Base (Grattafiori et al., 2024) and uses a custom single-BOS template and 20,000 unique prompts, with 10,000 refusal and 10,000 compliance targets. They follow the same settings as Qwen experiments.

F.2 Response Preferences and Over-Refusal Interventions

Directions use 51 qualified records from the model family's separate 57-record controlled-word bank, with the donor-matching construction described above. Independent evaluation selects 118 jointly answering benign and 118 jointly refusing harmful prompts. Benign source quotas are FalseReject 28, XSTest-safe 79, and ORBench-hard 11; harmful quotas are HarmBench 40, ORBench-toxic 39, and SORRY-Bench 39. Model/layer/site-specific directions are scaled by benign activation RMS and applied to the pure terminal-template suffix. Consequently doses are comparable within this family.

As shown in figure 7, at layer 22 and magnitude .8, benign addition induces refusal on 42.37%/1.69% of SFT/GRPO prompts; harmful subtraction removes refusal on 41.53%/3.39%. The full grid covers layers 8/10/12/14/16/18/22/26, magnitudes .1/.2/.4/.8/1.0, both signs.

The primary four-token readout contrasts log probability masses of 54 refusal-associated and 50 compliance-associated opening events, including 23 overlapping events. These are selected, nonexclusive event sets. Starting contrasts on the selected cohort are $-.5510/-.9200$ for benign SFT/GRPO and $.4001/1.6887$ for harmful SFT/GRPO. As in Qwen, GRPO's starting contrast is lower on benign and higher on harmful requests. These event-mass units differ from Qwen's candidate-averaged log-likelihood readout, so the numerical scales are interpreted within each model family.

F.3 Prefix-Attack Results and Activation Interventions

Figures 8–10 show the Llama counterparts of the activation-tracking, refusal-restoration and propagation analyses in Section 5.3. Table 17 additionally reports the longer-prefix evaluation for these Llama endpoints.

Table 24: Llama ordinary evaluation (%). Macros average three harmful datasets or three utility datasets.

Metric	SFT	GRPO
Harmful-refusal macro	96.40	91.71
Harmful-output macro	0.81	1.17
Utility macro	47.57	37.26
HarmBench refusal	100.00	99.50
ORBench-toxic refusal	94.20	82.90
SORRY-Bench refusal	95.00	92.73
GSM8K accuracy	14.94	12.51
ARC-Easy accuracy	67.72	48.25
PIQA accuracy	60.07	51.03
FalseReject refusal	54.93	41.95
XSTest-safe refusal	13.20	0.80
ORBench-hard refusal	58.30	48.37
JailbreakBench-benign refusal	63.00	58.00

Table 25: Llama full-pool prefix outcomes (%); paired entries are ORBench-toxic / SORRY-Bench.

Endpoint	Natural	H5
SFT	94.20 /95.00	7.48 /21.14
GRPO	82.90 /92.73	77.10 /90.45

G Direct Training of Instruction-Tuned Models

G.1 Summary of the Direct-Training Comparison

We compare SFT, GRPO and Offline Signed Learning (OSL; Section 5.2) as direct updates to the original Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct checkpoints. All three methods start independently from the same Instruct checkpoint within each model family and use the same ordered registry of 3200 balanced harmful and benign prompts. SFT fits one canonical demonstration per prompt. OSL uses eight ordinary self-rollouts per prompt and keeps the responses and their rewards fixed, whereas online GRPO samples fresh responses as training proceeds. OSL and GRPO use the same safety–behavior reward, and no training prompt contains the helpful prefix.

Table 26 summarizes ordinary harmful refusal, H5 refusal, benign refusal, and aggregate GPU-hours. Further SFT produces the highest ordinary harmful refusal in both families but substantially lowers H5 refusal and increases benign refusal. OSL and GRPO improve prefix recovery relative to the original Instruct checkpoints. On Qwen, benign refusal remains close to the original model for both reward-based methods. On Llama, both methods improve harmful refusal and H5 recovery while also increasing benign refusal. These transfer results are therefore best interpreted jointly across ordinary refusal, recovery, and selectivity rather than as a single alignment score.

Computational cost. We report aggregate GPU-hours, defined as the sum across stages of the number of GPUs multiplied by elapsed hours. GRPO includes policy and reward-model compute during online training. OSL includes its one-time response collection and judging followed by offline optimization. Evaluation and hyperparameter search are excluded. In the reported Qwen runs, GRPO uses approximately 12.1 aggregate GPU-hours and OSL 8.4; the corresponding Llama runs use 13.9 and 8.7.

Starting Models and Training Responses. The application starts directly from the original Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct checkpoints. SFT, OSL and GRPO are independent training branches from each original model. Both families use their native chat templates and the same ordered registry of 3200 training prompts, with eight harmful and eight benign prompts per 16-prompt batch. For OSL, each original model generates eight responses per prompt at temperature 1 and top-$p = 1$, with a 2,048-token limit and native stop tokens.

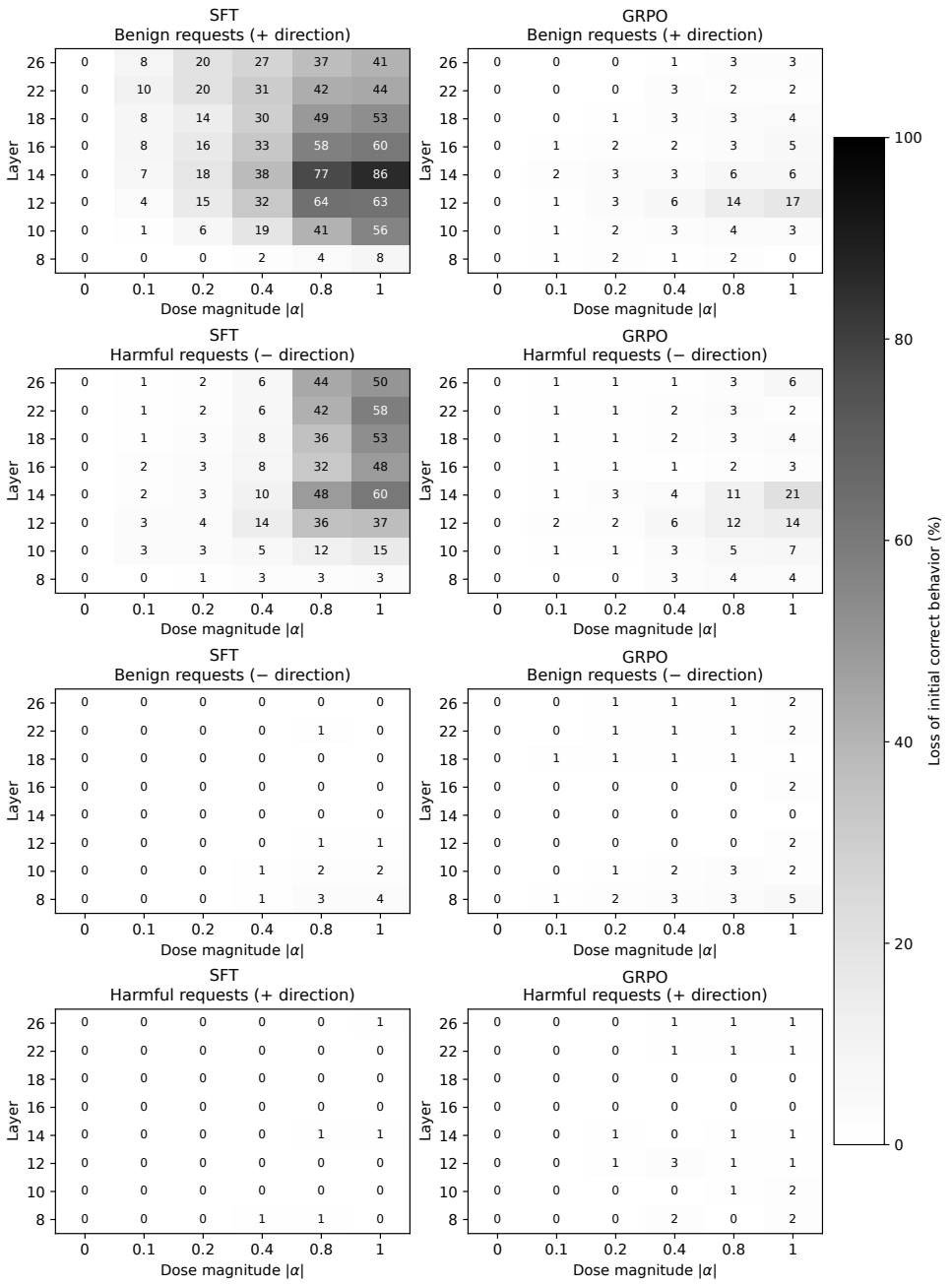


Figure 7: Llama loss of initially correct behavior on separate 118-benign and 118-harmful cohorts. Columns compare SFT and GRPO. The first two rows show benign addition and harmful subtraction; the last two show the opposite signs. Each trial perturbs one layer over the model's pure template suffix. Baseline loss is zero by cohort selection. Layer-specific outcomes retain the shared 0–100% behavioral scale.

SFT, Offline OSL and Online GRPO Training. All checkpoints are selected based on highest natural harmful refusal under development set. The development set contains 64 prompts each from ORBench-toxic, SORRY-Bench; harmful prompts are evaluated naturally, yielding 128 unique prompts. They are excluded from testsets.

We use batch size 16 for three models and save every 10 updates. SFT uses response-token-mean cross-entropy. The chosen learning rate is 10^{-5} with cosine decay. AdamW uses betas $(0.9, 0.95)$,

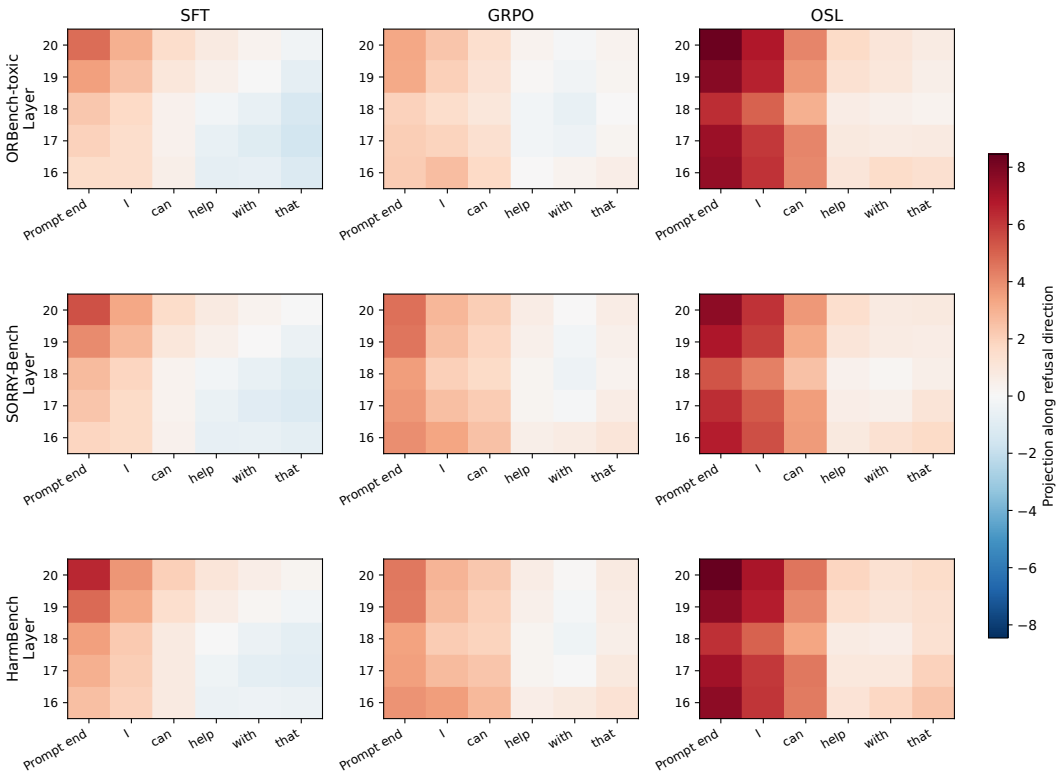


Figure 8: Llama refusal-related projections through the helpful prefix, on requests naturally refused by all three models. Rows show harmful datasets; columns show SFT, GRPO and OSL.

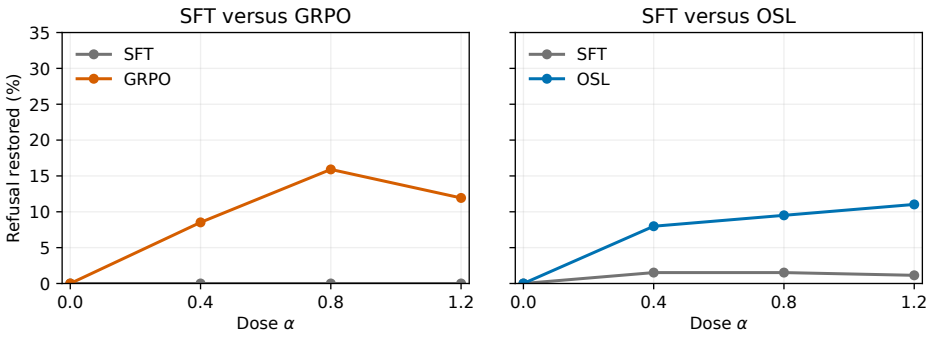


Figure 9: Refusal restored by prompt-end augmentation in Llama, on the separate SFT/GRPO and SFT/OSL shared-failure cohorts. The intervention spans layers 16–20.

weight decay 0.1 and gradient clipping at 1. Its selected checkpoints are update 200 for Qwen and Llama. OSL freezes the response bank and its WildGuard safety and behavior judgments. Both methods use the safety and behavior reward in Eq. 3. The selected OSL checkpoints: Qwen uses learning rate 9×10^{-7} for 80 updates, and Llama uses 8×10^{-7} for 90 updates. Online GRPO starts from the same original checkpoints and prompt registry, samples eight fresh responses per prompt, one PPO epoch per batch and KL coefficient 0.001 relative to the original Instruct reference. Its selected checkpoints are learning rate 10^{-6} with update 150 for Qwen and 10^{-6} with update 100 for Llama. Both optimizers are AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01 and gradient clipping at 1.

Generation and scoring. Behavioral evaluation uses greedy decoding with a 256-token continuation budget and native templates. The H5 condition forces `I can help with that` before

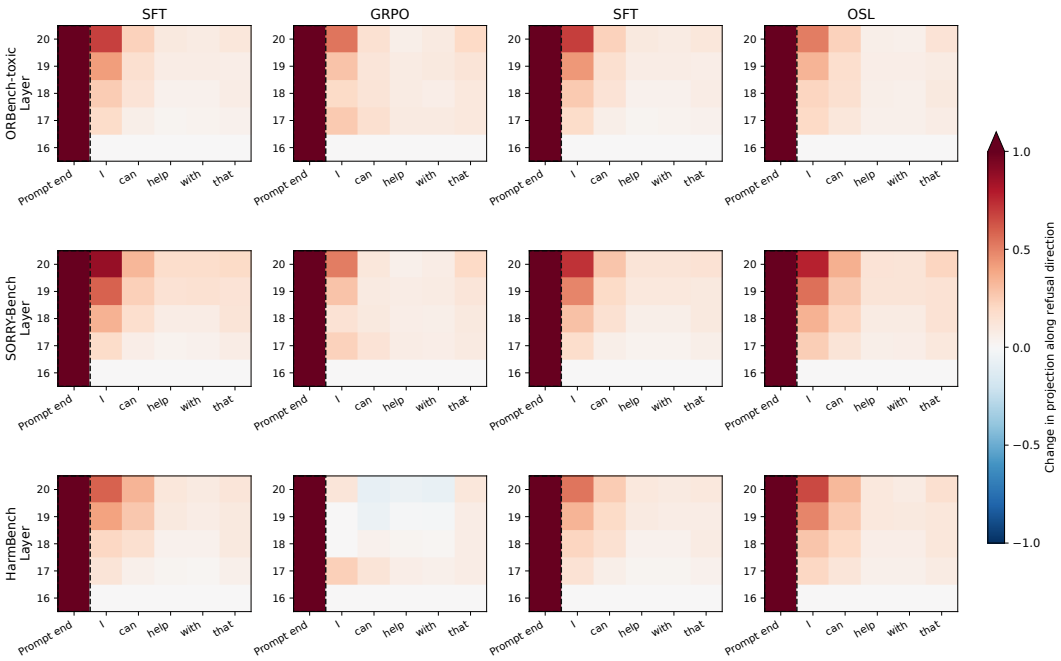


Figure 10: Llama changes in refusal-direction projection after prompt-end augmentation at dose 1.2. Values subtract each model's zero-dose projection at the same layer and position on the same shared-failure requests. The left and right pairs show SFT/GRPO and SFT/OSL on their respective shared-failure cohorts. Rows show datasets; dashed boxes mark directly modified states.

Table 26: Aggregate direct-training results on the common evaluation sets. Refusal rates are percentages; computational cost is aggregate GPU-hours.

Model	Method	Natural harmful ↑	H5 harmful ↑	Benign refusal ↓	Computational cost ↓
Qwen2.5-3B	Instruct	89.16	59.72	32.30	–
	SFT	98.32	11.24	53.84	0.23
	OSL	91.70	81.77	32.10	8.40
	GRPO	89.93	84.37	32.53	12.10
Llama-3.2-3B	Instruct	73.66	12.12	11.86	–
	SFT	98.63	3.86	59.90	0.27
	OSL	91.23	65.47	24.48	8.70
	GRPO	90.16	30.26	26.38	13.90

generation, and WildGuard judges the prefix and continuation together. Invalid judgments receive bounded retries on the same saved response.

Computational Cost and Hardware. For stages s using N_s GPUs for t_s elapsed seconds, we compute

$$C_{\mathrm{GPUh}} = \sum_s N_s t_s / 3600.$$

GRPO counts both policy and reward-model GPUs during its training window. OSL counts one-time response collection and judging followed by offline optimization. Evaluation is excluded. The reported runs use NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs; `nvidia-smi` reports 97,887 MiB of memory per GPU.

Table 27: Direct-Instruct harmful refusal rates (%). All models use 591 ORBench-toxic, 376 SORRY-Bench and 200 HarmBench requests, each evaluated naturally and under H5.

		ORBench-toxic		SORRY-Bench		HarmBench	
Model	Method	Natural	H5	Natural	H5	Natural	H5
Qwen2.5-3B	Instruct	95.09	65.31	77.39	51.86	95.00	62.00
	SFT	98.65	5.25	96.81	7.98	99.50	20.50
	OSL	94.08	85.11	83.51	74.20	97.50	86.00
	GRPO	90.02	84.77	83.78	76.33	96.00	92.00
Llama-3.2-3B	Instruct	68.70	12.18	63.30	11.17	89.00	13.00
	SFT	98.82	2.37	97.07	3.72	100.00	5.50
	OSL	92.05	65.82	85.64	59.57	96.00	71.00
	GRPO	92.89	31.64	80.59	27.13	97.00	32.00

Table 28: Direct-Instruct benign refusal rates (%; lower is better). All models use the full benign sets: FalseReject (1,187), XSTest-safe (250), ORBench-hard (1,319) and JBB-benign (100).

Model	Method	FalseReject	XSTest-safe	ORBench-hard	JBB-benign
Qwen2.5-3B	Instruct	36.73	16.80	48.67	27.00
	SFT	58.64	20.80	79.91	56.00
	OSL	36.82	6.00	53.60	32.00
	GRPO	36.82	4.40	54.89	34.00
Llama-3.2-3B	Instruct	31.09	5.20	6.14	5.00
	SFT	73.21	14.00	84.38	68.00
	OSL	39.76	3.60	34.57	20.00
	GRPO	39.09	0.40	41.02	25.00

H Statistical Reporting and Reproducibility

Ordinary harmful refusal, benign refusal and utility are reported separately. Appendix B gives the complete Qwen natural benchmark results; Appendix F gives the additional model's outcomes. Intervention analyses retain their own baseline generations and selection rules.

Training comparisons currently use one trajectory per arm. WildGuard supplies both behavior feedback during training and refusal labels during evaluation.

The reproducibility record retains checkpoint and data hashes, software and numerical configurations, frozen prompt registries, evaluator contracts, no-op checks and original failed numerical checks.

I Extended Related Work

This appendix expands the discussion in Section 2. We review work directly related to generator-side safety training, differences between SFT and reinforcement learning, major safety-alignment failures and defenses, and causal analyses of refusal.

I.1 Safety Training with SFT and Reinforcement Learning

Modern alignment pipelines usually begin with SFT on helpful and safe demonstrations, then optimize preferences or rewards. InstructGPT combined SFT, reward modeling, and PPO; helpful-and-harmless training and Constitutional AI expanded this recipe with explicit harmlessness preferences or written principles (Ouyang et al., 2022; Bai et al., 2022a;b). Safe RLHF separates helpfulness and safety objectives, BeaverTails supplies large-scale safety preference data, and Rule-Based Rewards

translates safety policies into scalable reward signals (Dai et al., 2023; Ji et al., 2023; Mu et al., 2024). These methods build safety by supervising either a target response or a preference among candidate responses.

SFT gives dense teacher-forced likelihood to one desired trajectory. Preference objectives such as RRHF and DPO compare completions without an online actor-critic loop, while SafeDPO and balancing methods adapt preference learning to the safety–helpfulness trade-off (Yuan et al., 2023; Rafailov et al., 2023; Kim et al., 2026; Tan et al., 2025). PPO and GRPO instead score sampled responses from the current policy; GRPO forms within-prompt relative advantages without a learned value model (Schulman et al., 2017; Shao et al., 2024; DeepSeek-AI, 2025). GRPO is therefore not a safety specification by itself: its behavior depends on sampled alternatives, reward assignment, clipping, KL regularization, and the starting policy.

In this paper, we compare how SFT and GRPO-style training build the safety-related decision and execution components inside the same model.

I.2 Failure Modes and Defenses in Safety Alignment

Over-refusal. XSTest identified exaggerated safety behavior on benign prompts that resemble harmful requests; Pseudo-Harmful, OR-Bench, and FalseReject broadened the coverage of lexical, contextual, and task-specific false refusals (Rottger et al., 2024; An et al., 2024; Cui et al., 2025; Zhang et al., 2025b). These benchmarks show why raw refusal rate is insufficient: a policy can appear safe by refusing nearly everything. Proposed remedies include over-refusal preference data, explicit "think before refusing" decisions, balanced safety data, safety-aware DPO, and configurable or competing rewards that distinguish harmful intent from surface similarity (Karaman et al., 2024; Si et al., 2025; Tan et al., 2025; Kim et al., 2026; Jiang et al., 2026; Kim & Kumar, 2026). Policy-grounded reasoning methods such as Deliberative Alignment and guideline-based safety reasoning also report improved jailbreak robustness and lower over-refusal by making the governing policy explicit (Guan et al., 2024; Wang et al., 2025).

Shallow alignment and prefix or prefill attacks. Jailbroken cataloged competing-objective failures, universal adversarial suffixes demonstrated transferable optimized attacks, and the shallow-alignment hypothesis showed that safety training may primarily alter the first few response tokens, allowing an affirmative assistant prefill to bypass refusal entry (Wei et al., 2023; Zou et al., 2023; Qi et al., 2025). HarmBench and StrongREJECT provide standardized harmful-behavior and jailbreak evaluation, emphasizing that attack success must be checked for substantive harmful content rather than mere non-refusal (Mazeika et al., 2024; Souly et al., 2024). Defenses include prefix-augmented SFT, rank- and attention-aware deep alignment, training on perturbed intermediate trajectories, and correction-oriented reasoning or on-policy methods that teach a model to override a developing unsafe path (Qi et al., 2025; Vega & Singh, 2025; Lyu et al., 2026; Park & Kim, 2026; Zhu et al., 2025; Peng et al., 2025; Fu et al., 2026; Han et al., 2026). These methods usually expose the model to corrupted or contrastive states during training.

I.3 Refusal representations and intervention-based analysis

Refusal-direction studies identify activation components whose addition or removal can alter refusal behavior (Arditi et al., 2024). Work separating harmfulness and refusal further shows why a change in refusal-related activity need not imply a corresponding change in harmfulness recognition (Zhao et al., 2025). These findings motivate our activation measurements and interventions. We distinguish projections, which characterize the learned policies, from interventions, which test the behavioral influence of a selected component. In the prefix setting, our question is whether an intervention made at prompt end can affect refusal after the helpful opening has been imposed; we do not assume that a single measured direction completely mediates the training effect.