HARM RECOGNITION TO REFUSAL GAP: WHY SMALL LLMS ARE UNSAFE?

Anonymous authors
Paper under double-blind review

ABSTRACT

The growing deployment of small language models makes their vulnerability to jailbreaks an important safety concern. A common belief is that small models are less robust to jailbreaks because they have weaker representation, reasoning, and instruction-following abilities, and thus fail to recognize disguised harmful intent. However, we find that small models can still clearly separate harmful from benign inputs in their internal representations. Their failures arise because this harmfulness-related signal does not reliably translate into refusal behavior. Using existing harmfulness and refusal directions, we show that prompts with strong harmfulness-related activations can still elicit compliant or harmful responses. This suggests a bottleneck in the internal conversion from harmfulness recognition to refusal execution, which we call the recognition-to-refusal gap. Across model scales, safety differences are more closely associated with the coupling between these two signals than with refusal strength alone. Motivated by this finding, we propose Coupling-Aware Direct Preference Optimization (CA-DPO). CA-DPO focuses training on jailbreak cases where harmfulness recognition does not lead to refusal, while retaining benign examples to preserve helpfulness and limit over-refusal. Experimental results show that CA-DPO improves model safety while mitigating over-refusal, suggesting that optimizing the recognition-to-refusal pathway can improve safety while preserving useful behavior in safety-critical deployment.

1 INTRODUCTION

Small language models are increasingly deployed in local assistants, edge devices, and privacy-sensitive applications, where compact models provide direct on-device interfaces under tight memory, latency, and compute constraints (Liu et al., 2024b; Abdin et al., 2024; Gunter et al., 2024). Unlike large models, which often rely on cloud infrastructure, small models may serve as the direct interface between users and applications, interpreting requests, making decisions, and generating responses locally. In settings that cannot rely on cloud-side moderation, the model must handle potentially unsafe user inputs under the same resource constraints that motivate local inference (Yi et al., 2025; Wang et al., 2025b). Small-model safety is therefore an issue in its own right, beyond the lower cost of deployment, and is essential to the reliable use of local AI systems.

Existing studies generally attribute the higher rate of safety failures in small models to weaker representation, reasoning, and instruction-following abilities(Yi et al., 2025; Wang et al., 2025b). These limitations also make it harder to recognize harmful intent hidden through paraphrasing, encoding, or indirect wording(Wang et al., 2025b). Although reasonable, this explanation largely treats small-model safety as a weaker form of large-model safety: limited knowledge and understanding make harmful responses more likely. It leaves a more specific question unanswered: when a small model fails on a jailbreak prompt, does it fail to recognize the risk, or does it form an internal representation of that risk but fail to act on it by refusing?

Harm recognition and refusal execution are not necessarily the same process. Prior mechanistic studies suggest that refusal behavior may be associated with a specific direction in the model's internal representation space (Arditi et al., 2024), and that the encoding of input harmfulness may be separate from the final refusal behavior (Zhao et al., 2026). A safety response can thus be understood as involving at least two related but distinct stages: recognizing risk in an input and translating that recognition into a refusal. Treating these stages as a single capability can lead us to attribute unsafe behavior solely to a failure to understand harmful intent, overlooking another important mode of failure.

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 whether jailbreak failures in small models involve an internal mechanism beyond an overall weakening of large-model safety capabilities. Our experiments show that small models retain harmfulness-related information. Even at smaller scales, their internal representations distinguish jailbreak prompts, normal safe inputs, and over-refusal inputs. These findings indicate that the failures cannot be explained simply by an inability to form risk-related representations. Yet the presence or strength of a risk-related signal alone does not determine whether a model refuses. A prompt can activate a strong harmfulness representation while still eliciting a compliant or harmful response. Recognizing risk, in other words, does not ensure refusal.

These results show that insufficient harmfulness recognition alone cannot explain safety failures in small models. A key bottleneck arises when risk-related information is present in the model's internal representations but does not reliably guide subsequent generation and trigger refusal. We call this disconnect between harmfulness recognition and refusal execution the *recognition-to-refusal gap*. To characterize it, we model the two signals separately and examine their relationship through prompt-level mismatch and representation-level coupling. This framework distinguishes a failure to recognize risk from a failure to act on recognized risk. We do not claim that all jailbreak failures have the same cause. Instead, we identify a systematic failure in safety expression that explains why small models can remain vulnerable even when harmfulness-related information is available internally.

We further show that larger models may be safer because recognized risk is more reliably converted into refusal. Larger models may have stronger risk representations and more stable refusal behavior, but the differences across scales extend beyond a stronger refusal prior. Our results show that coupling between harmfulness recognition and refusal execution improves as model scale increases, making recognized risk more likely to lead to an appropriate refusal. For small models, the main bottleneck in some cases is the inability to reliably translate existing risk recognition into refusal, rather than a complete lack of understanding of what is dangerous. Safety failures in small models should therefore be understood as problems in safety expression that may involve a different bottleneck, beyond an overall weakening of large-model safety.

Based on these findings, we propose *Coupling-Aware Direct Preference Optimization (CA-DPO)* to repair the recognition-to-refusal pathway. CA-DPO trains on jailbreak cases where harmfulness recognition fails to trigger refusal, while using benign and over-refusal dates to keep the model useful on safe inputs. This supervision strengthens the missing link from recognized harm to refusal and reduces the risk that safety tuning becomes a broad refusal bias. These design choices address distinct concerns documented in prior work: downstream fine-tuning can weaken safety alignment, excessive safety tuning can induce over-refusal, and preference optimization can reduce the likelihood of desired responses (Qi et al., 2025; 2024; Bianchi et al., 2024; Ren & Sutherland, 2025). Over-refusal should therefore be evaluated separately from harmful-request refusal (Röttger et al., 2024). Experiments across model scales and settings verify that CA-DPO improves model safety and reduces over-refusal without substantially reducing model utility. Our contributions are as follows:

- We examine jailbreak failures in small models from a recognition-to-refusal perspective. We show that these failures may arise from a specific bottleneck in translating harmfulness recognition into refusal execution, rather than simply reflecting weaker safety capabilities than those of larger models.
- We distinguish harmfulness recognition from refusal execution and develop a diagnostic framework at the representation level. We find that changes in recognition-to-refusal coupling better explain safety differences across model scales than changes in refusal strength alone.
- We introduce a mechanism-guided strategy for constructing targeted safety preference data. The strategy selects jailbreak prompts with strong harmfulness recognition, weak refusal execution, and a large recognition-to-refusal mismatch.
- We propose CA-DPO, combining mechanism-guided preference data with coupling-aware optimization to repair the recognition-to-refusal pathway while controlling over-refusal and response collapse.

2 Related Work

Small language models. Small language models (SLMs) are increasingly studied for code generation, reasoning, long-context processing, and local inference (Wang et al., 2025a; Corradini

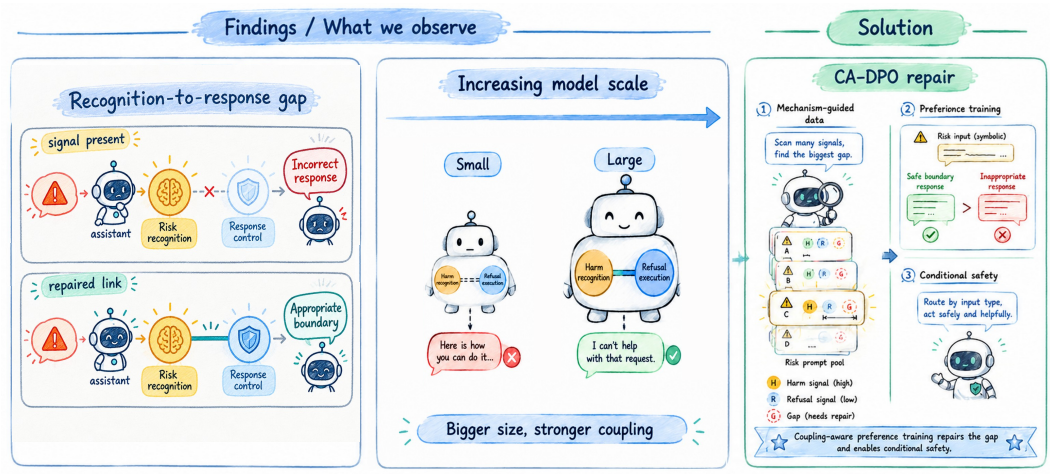


Figure 1: Overview of the recognition-to-refusal gap in small language models and the proposed CA-DPO repair. The left panel illustrates that harmfulness recognition may be present while response control fails, producing an inappropriate response; repairing the coupling between the two signals enables an appropriate refusal. The middle panel shows how this relationship changes with model scale: larger models exhibit stronger coupling and more reliable refusal, whereas smaller models may recognize harmfulness yet still comply. The right panel summarizes CA-DPO, which selects prompts with high harmfulness recognition, weak refusal execution, and a large recognition-to-refusal gap, then uses conditional preference training to preserve helpful behavior on benign inputs while enforcing safer responses to harmful requests.

et al., 2025; Lu et al., 2024). Model families such as Llama, Qwen, and Gemma have also released compact variants with improving instruction-following, reasoning, and multimodal capabilities (Grattafiori et al., 2024; Yang et al., 2025; Team et al., 2025). Although no universal size threshold exists, we use models with at most 7B parameters as the operational scope, following recent survey and evaluation practice (Corradini et al., 2025; Wang et al., 2025c). Recent studies find that these models remain vulnerable to jailbreaks (Wang et al., 2025c); meanwhile, EASE, distillation studies, and SSAH show that safety alignment must balance robustness, inference cost, and utility (Shi et al., 2026; Padarha et al., 2026; Li & Kim, 2026; Huang et al., 2025).

Representation-level safety and refusal directions. Representation-level methods analyze safety behavior through internal activations(Zhou et al., 2024; Wollschläger et al., 2025). Representation Engineering provides a framework for reading and controlling high-level concepts in activation space (Zou et al., 2023a). Arditi et al. show that refusal in aligned chat models is often mediated by an approximately one-dimensional direction, while Zhao et al. distinguish harmfulness representations from refusal representations and show that jailbreaks can weaken refusal without removing harmfulness recognition (Arditi et al., 2024; Zhao et al., 2026). Recent analyses further suggest that a single refusal direction may conflate harm detection with the execution of refusal (Zhang & Sun, 2026).Related defenses reroute harmful representations, steer safety activations, or explicitly couple harmfulness and refusal directions, as in Circuit Breakers, SCANS, and HARC (Zou et al., 2024; Cao et al., 2025; Chua et al., 2026).

Jailbreak defenses and over-refusal. Safety alignment must address both harmful responses under jailbreak attacks and unnecessary refusals of benign requests. Universal adversarial suffixes and AutoDAN expose jailbreak vulnerabilities, while JailbreakBench standardizes robustness evaluation (Zou et al., 2023b; Liu et al., 2024a; Chao et al., 2024). Existing defenses include training-based methods such as SFT (Ouyang et al., 2022), RLHF (Ouyang et al., 2022; Bai et al., 2022), DPO (Rafailov et al., 2023), and goal prioritization (Zhang et al., 2024), as well as inference-time methods such as SafeDecoding (Xu et al., 2024) and SmoothLLM (Robey et al., 2023). Other approaches steer activations or reduce excessive sensitivity to harmful-looking wording (Cao et al., 2025; Shi et al., 2024). Meanwhile, XSTest, OR-Bench, FalseReject, and WildGuard show that safety mechanisms can over-refuse benign prompts and motivate evaluating jailbreak robustness together with refusal quality (Röttger et al., 2024; Cui et al., 2024; Zhang et al., 2025; Han et al., 2024).

3 DO SMALL MODELS FAIL DIFFERENTLY?

3.1 EXPERIMENTAL SETUP AND SIGNAL DEFINITIONS

Harmfulness and refusal directions. Following prior work, we use separate directions in representation space to characterize harmfulness and refusal (Arditi et al., 2024; Zhao et al., 2026). For a model f_θ, let $h_{\theta,\ell}(x) \in \mathbb{R}^d$ denote the hidden state at a specified position in layer ℓ. We extract harmfulness-related representations at the instruction-side position and refusal-related representations at the post-instruction position. The mean hidden states for harmful and clean instructions are denoted by $\mu^{(\ell)}_{\text{harm}}$ and $\mu^{(\ell)}_{\text{clean}}$, respectively; those for refused and non-refused samples are denoted by $\mu^{(\ell)}_{\text{ref}}$ and $\mu^{(\ell)}_{\text{nonref}}$. We normalize the differences between these means to obtain the two directions:

$$v^{(\ell)}_{\text{harm}} = \frac{\mu^{(\ell)}_{\text{harm}} - \mu^{(\ell)}_{\text{clean}}}{\left\|\mu^{(\ell)}_{\text{harm}} - \mu^{(\ell)}_{\text{clean}}\right\|_2}, \qquad r^{(\ell)}_{\text{ref}} = \frac{\mu^{(\ell)}_{\text{ref}} - \mu^{(\ell)}_{\text{nonref}}}{\left\|\mu^{(\ell)}_{\text{ref}} - \mu^{(\ell)}_{\text{nonref}}\right\|_2}. \tag{1}$$

We measure directional coupling between harmfulness recognition and refusal execution using cosine similarity:

$$\text{Cos}_\ell = \cos\left(v^{(\ell)}_{\text{harm}}, r^{(\ell)}_{\text{ref}}\right) = \frac{\left\langle v^{(\ell)}_{\text{harm}}, r^{(\ell)}_{\text{ref}} \right\rangle}{\left\|v^{(\ell)}_{\text{harm}}\right\|_2 \left\|r^{(\ell)}_{\text{ref}}\right\|_2}. \tag{2}$$

Both directions have unit norm, so Cos_ℓ equals their inner product. Larger values indicate closer alignment at layer ℓ.

Given an input sample x, we project its hidden states onto the two directions:

$$\widetilde{H}_\ell(x) = \left\langle h^{\text{inst}}_{\theta,\ell}(x), v^{(\ell)}_{\text{harm}} \right\rangle, \qquad \widetilde{R}_\ell(x) = \left\langle h^{\text{post}}_{\theta,\ell}(x), r^{(\ell)}_{\text{ref}} \right\rangle. \tag{3}$$

The raw projection scores $\widetilde{H}_\ell(x)$ and $\widetilde{R}_\ell(x)$ may differ in scale. We therefore standardize each score using its empirical mean and standard deviation on the corresponding calibration set:

$$H_\ell(x) = \frac{\widetilde{H}_\ell(x) - \mu_{H,\ell}}{\sigma_{H,\ell}}, \qquad R_\ell(x) = \frac{\widetilde{R}_\ell(x) - \mu_{R,\ell}}{\sigma_{R,\ell}}. \tag{4}$$

The standardized scores $H_\ell(x)$ and $R_\ell(x)$ measure the relative strength of a sample's harmfulness-related and refusal-related representations. These are projection scores, not probabilities. We define their difference as

$$G_\ell(x) = H_\ell(x) - R_\ell(x), \tag{5}$$

where $G_\ell(x)$ measures the gap between the two standardized signals. Larger values indicate a stronger harmfulness signal relative to the refusal signal.

Data and models. We analyze the Qwen3.5 and Llama3.2 model families, which offer strong general capabilities and instruction following at relatively small model sizes. These capabilities make them suitable for studying safety behavior in small models. We use each model's native tokenizer and chat template and estimate the representation directions separately for each model. For jailbreak data, we follow the competing-objectives formulation of Wei et al. (Wei et al., 2023) when selecting attack samples; full details are provided in Appendix A.

3.2 LAYER-WISE REPRESENTATION VISUALIZATION AND ANALYSIS

If jailbreak failures in small models arose mainly from an inability to form safety-related representations, harmful and jailbreak inputs would be expected to overlap substantially with clean inputs in representation space, or at least show much weaker separation than in larger models. To examine this possibility, we use PCA to visualize intermediate-layer representations in Qwen3.5 models of different sizes and compare the separation of harmful inputs from the other sample groups.

In Fig. 2, harmful and jailbreak inputs are already separated from clean inputs in early layers of both small and large models. This suggests that jailbreak failures cannot be attributed solely to missing safety-related input representations. It remains unclear why these representations do not reliably translate into refusal behavior.

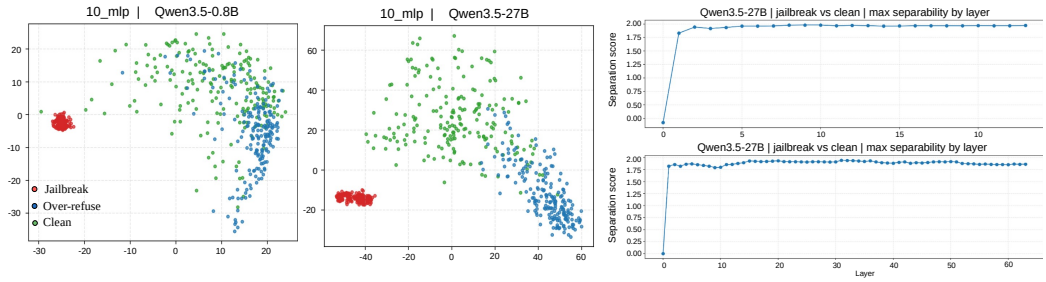


Figure 2: Layer-wise PCA visualization of jailbreak, over-refusal, and clean prompts across Qwen3.5 model sizes. We compare intermediate-layer representations to test whether safety failures arise from missing harmfulness-related input representations. The early-layer separation of jailbreak prompts from clean prompts suggests that smaller models do not fail solely because they lack safety-related representations.

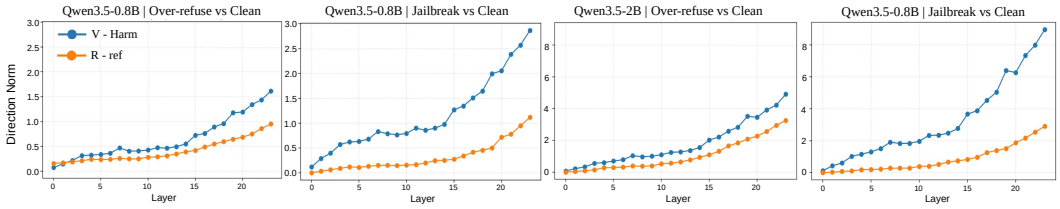


Figure 4: Harmfulness-related and refusal-related activation strengths across jailbreak and over-refusal conditions. The figure compares the input-side harmfulness signal with the refusal-related signal across safety behaviors.

3.3 Scale-Dependent Imbalance Between Recognition and Refusal

Sec. 3.2 shows that models of different sizes separate jailbreak inputs from clean inputs in their internal representations. To examine how model size affects the relation between harmfulness recognition and refusal, we compare the relative strengths of the harmfulness-related and refusal-related signals. For each model, we average each signal's strength across layers before computing the ratio of harmfulness to refusal. This comparison examines whether an encoded harmfulness signal is accompanied by a correspondingly strong refusal signal.

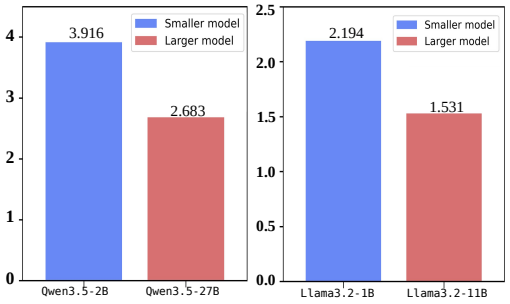


Figure 3: Relative harmfulness-related and refusal-related activation across representative different models. We compute the harmfulness-to-refusal relative signal-strength ratio across layers; its decrease with scale is consistent with the hypothesis that larger models more reliably convert recognized risk into refusal.

As shown in Fig. 3, the harmfulness-related signal is stronger than the refusal-related signal across models, but their ratio decreases as model size increases. Small models thus exhibit harmfulness-related representations alongside relatively weak refusal signals. In larger models, the refusal signal is stronger relative to the harmfulness signal, reducing this imbalance.

These results suggest that larger models may be better at translating an existing risk representation into refusal behavior. Small models may recognize some risk yet fail to use that information consistently during generation. Model size may therefore affect the reliability of recognition-to-refusal conversion, beyond whether harmfulness is represented at all.

This interpretation requires further evidence from the representations. This raises a more diagnostic question: when small models recognize harmfulness, do their refusal directions become coupled with that recognition strongly enough to separate refused from non-refused jailbreaks?

3.4 Recognition-Refusal Signals Across Safety Behaviors

We apply the same calculation procedure to jailbreak and over-refusal data conditions and construct the input-side and refusal-related directions separately. As shown in Fig. 4, the strength of the input-side direction is substantially lower under the over-refusal condition than under the jailbreak con-

dition, indicating a weaker difference in harmfulness-related representations between over-refusal and clean samples. However, the strengths of the refusal-related directions are similar across the two conditions, and their final refusal rates are also broadly comparable. This indicates that over-refusal can produce a refusal-related signal comparable to that of jailbreak under a weaker input-side risk-related representation. Details of the construction are provided in Appendix A.

This result is consistent with two distinct safety failure mechanisms. For over-refusal samples, the model does not form a strong harmfulness-related representation, but some words or expressions in the input may share surface associations with harmful contexts in the pretraining or safety-alignment data. The model may therefore treat them as cues for refusal and trigger an unnecessary refusal. For jailbreak samples, the model has already formed a stronger harmfulness-related signal, but the strength of the refusal-related direction does not increase accordingly. This means that the model can recognize the risk in the input, yet fails to reliably convert that recognition into refusal behavior.

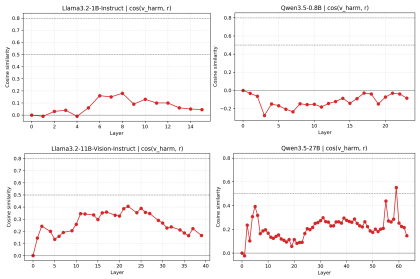


Figure 5: Layer-wise coupling between harmfulness-related and refusal-related directions across different models. We measure cosine similarity across layers and model sizes to assess whether recognized risk aligns with refusal execution, motivating sample-level comparisons within each matched attack family.

Jailbreak and over-refusal are therefore not the same failure at different intensities. The former reflects stronger input-side risk recognition that is not sufficiently transferred to the refusal-related state, whereas the latter reflects a weaker input-side risk representation that can still trigger a comparable refusal response. This result suggests that model safety cannot be judged solely by the strength of the refusal-related signal; it is also necessary to examine whether refusal behavior is grounded in corresponding harmfulness recognition. We further analyze the coupling between the two directions and its relationship with model safety.

3.5 Recognition-to-Refusal Coupling and Behavioral Differences

Signal strength alone does not establish a link between harmfulness recognition and refusal. Both signals may be present without the harmfulness representation influencing refusal behavior. We therefore examine the cosine similarity between v_{harm} and r_{ref} and compare jailbreak inputs that elicit refusal with those that do not.

Fig. 5 shows that coupling varies across model sizes. Within the same attack family, inputs that elicit refusal generally show stronger coupling than those that do not, as shown in Tab. 1. This pattern suggests that refusal depends on more than the strength of the harmfulness-related signal. Even when risk-related information is present, refusal appears more likely when that information aligns closely with the refusal-related representation.

Jailbreak failures may arise between harmfulness recognition and refusal execution. A model can represent harmfulness without reliably using it to refuse, producing high-H/low-R samples. Refused jailbreaks show coordinated activation of both signals, whereas non-refused inputs indicate a failure to convert recognition into refusal.

Table 1: Sample-level coupling for refused and non-refused jailbreaks. Higher cosine values for refused samples indicate that jailbreaks are more likely to be refused when harmfulness-related and refusal-related signals are better aligned.

Model	Split	cipher	h4rm3l	hand
Qwen3.5-0.8B	Refused	**0.069**	**-0.020**	**0.171**
Qwen3.5-0.8B	Non-refused	0.024	-0.030	0.115
Qwen3.5-2B	Refused	**0.009**	**0.099**	**-0.013**
Qwen3.5-2B	Non-refused	-0.008	0.067	-0.047

The recognition-to-refusal gap may therefore be a key mechanism behind safety failures in small models. This interpretation motivates our method, which targets high-H/low-R samples to improve the reliability with which recognized risk leads to refusal, rather than simply increasing the model's overall tendency to refuse.

4 CA-DPO: Mechanism-Guided Repair

CA-DPO combines mechanism-guided preference data with a coupling-aware objective. Its goal is to improve safety on harmful prompts while preserving the model's general response capability and reducing unnecessary refusals on benign prompts. The method addresses under-refusal

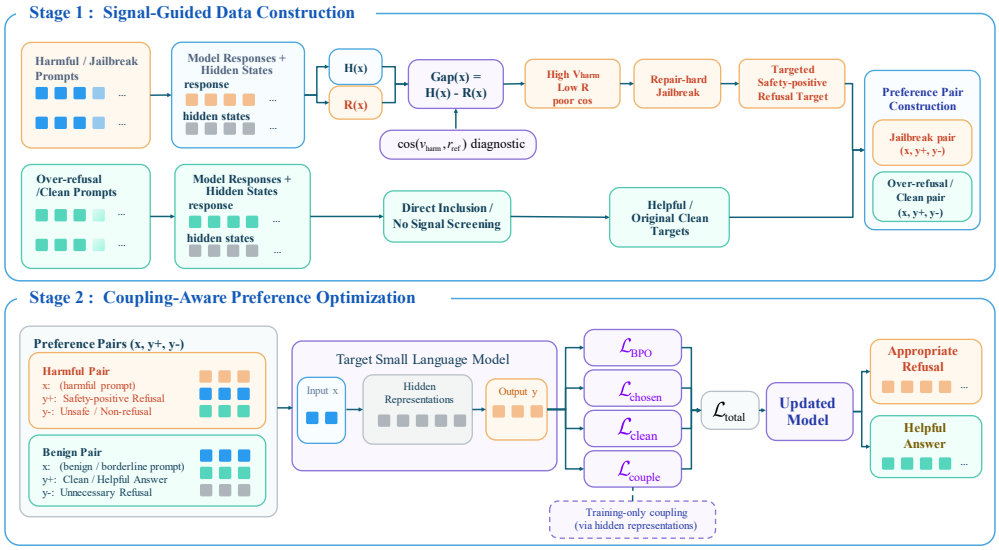


Figure 6: Overview of CA-DPO. Stage 1 constructs preference pairs by identifying jailbreak prompts with a large recognition-to-refusal gap and pairing benign or over-refusal prompts with helpful targets. Stage 2 optimizes the target small language model with preference learning, output stabilization, clean-behavior preservation, and a training-only hidden coupling loss.

on jailbreak prompts, over-refusal on safe prompts, response collapse during preference tuning, and weak coupling between harmfulness recognition and refusal execution.

The preference data comprise three prompt groups. We denote jailbreak, clean, and over-refusal examples by $\mathcal{D}_{\text{jb}}$, $\mathcal{D}_{\text{clean}}$, and $\mathcal{D}_{\text{or}}$, respectively. For a jailbreak prompt $x \in \mathcal{D}_{\text{jb}}$, we compute $H(x)$, $R(x)$, and $G(x) = H(x) - R(x)$, and select prompts with high harmfulness recognition, weak refusal execution, and positive mismatch. The preferred response y^+ is a safe refusal target, while the dispreferred response y^- is an unsafe or compliant response.

Clean and over-refusal examples preserve normal responses on safe prompts. The clean set $\mathcal{D}_{\text{clean}}$ contains benign prompts for which helpful responses should be maintained, while $\mathcal{D}_{\text{or}}$ contains safe prompts that are incorrectly refused by the original model. For both groups, the preferred response y^+ is a helpful answer and the dispreferred response y^- is an unnecessary refusal or a low-quality safety response. The three groups are used together for preference optimization, while the clean examples also provide the clean SFT anchor. Guided generation constructs target responses that specify desired behavior rather than imitate the model's original outputs.

The preference data are optimized with DPO. For harmful prompts, the model is trained to prefer safe refusal over unsafe compliance; for clean and over-refusal prompts, it is trained to prefer helpful answers over unnecessary refusals. CA-DPO uses DPO as its preference-learning backbone and adds auxiliary terms for response quality, and recognition-to-refusal coupling.

The main preference term aligns chosen and rejected responses under a reference-regularized objective. For each preference tuple (x, y^+, y^-), CA-DPO uses the standard DPO loss:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y^+,y^-)} \log \sigma \left(\beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)} \right] \right), \tag{6}$$

where π_θ is the trainable model, π_{ref} is the frozen reference model, and β controls the strength of the preference update. On jailbreak prompts, this term increases the relative likelihood of safe refusals over unsafe compliance. On benign and over-refusal prompts, it increases the relative likelihood of helpful answers over unnecessary refusals.

CA-DPO adds output anchors to reduce response collapse during small-model preference tuning. Small models can produce degenerate responses during aggressive safety tuning, including repeated templates, malformed outputs, or empty safety statements. To stabilize generation, CA-DPO adds chosen-response anchor and clean SFT anchor::

$$\mathcal{L}_{\text{chosen}} = -\mathbb{E}_{(x,y^+)} \log \pi_\theta(y^+ \mid x), \tag{7}$$

$$\mathcal{L}_{\text{clean}} = -\mathbb{E}_{(x,y_{\text{clean}})\in\mathcal{D}_{\text{clean}}} \log \pi_\theta(y_{\text{clean}} \mid x). \tag{8}$$

The chosen-response anchor keeps the model close to high-quality preferred responses, while the clean anchor preserves normal helpful behavior on benign prompts. These anchors are not added as generic regularizers; they address the observed failure that small models can improve refusal rate while degrading output quality.

The hidden coupling loss promotes refusal execution under high harmfulness recognition. The core coupling term operates on hidden representations. For harmful prompts, CA-DPO raises the refusal-execution score $R_\theta(x)$ under a high harmfulness recognition score $H(x)$:

$$\mathcal{L}_{\text{couple}} = \mathbb{E}_{x\in\mathcal{D}_{\text{jb}}} \left[\text{sg}(H(x)) \cdot (\text{sg}(H(x)) - R_\theta(x))_+^2 \right], \tag{9}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $(\cdot)_+$ denotes the positive part. This loss is not activation steering at inference time. It is a training-time regularizer that strengthens the pathway from harmfulness recognition to refusal execution. The stop-gradient operation is essential: $H(x)$ is used as a condition indicating that harmfulness recognition is present, not as a target to be amplified. Without this constraint, the model could reduce the loss by increasing harmfulness-sensitive activations globally, which would risk over-refusal. With stop-gradient, optimization focuses on the missing refusal-execution component.

The final objective combines preference learning, output stabilization, clean-behavior preservation, and hidden coupling. The complete CA-DPO objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DPO}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{chosen}} + \lambda_{\text{clean}}\mathcal{L}_{\text{clean}} + \lambda_{\text{hidden}}\mathcal{L}_{\text{couple}}. \tag{10}$$

During training, $H_\theta(x)$ uses the prompt-side representation at ℓ_h, while $R_\theta(x)$ uses the assistant-side representation at ℓ_r. We select ℓ_h and ℓ_r following Sec. 3 and keep them fixed during optimization. We apply the DPO objective to preference pairs from all three groups, use the clean SFT anchor on $\mathcal{D}_{\text{clean}}$, and apply the coupling loss only to jailbreak prompts in $\mathcal{D}_{\text{jb}}$. At inference time, CA-DPO requires no additional controller, activation intervention, or external classifier.

5 Experiments

5.1 Experimental Setup

This section largely follows the configuration used in Sec. 3, and further includes the Gemma3 family to evaluate the effectiveness of CA-DPO . The training and test pools are kept strictly disjoint. JB refuse, Clean refuse, and OR refuse are computed as the average refusal rates over the corresponding jailbreak, clean, and over-refusal test sets, respectively. The source composition, sampling rules, and splits for jailbreak, clean, and over-refusal data are provided in Appendix A.

For model responses, we define a refusal as any output that contains predefined refusal keywords (Zou et al., 2023a). Since smaller models have fewer parameters and are more prone to degenerate outputs after safety tuning, we additionally report collapse rate, which captures repeated templates, empty or uninformative answers, malformed formats, and abnormal response lengths. The detailed judging rules and implementation details are provided in Appendix B.

Because our goal is to handle jailbreak and over-refusal simultaneously, we compare CA-DPO with baselines under the same dual-objective setting. Safety SFT (Ouyang et al., 2022) and Vanilla DPO (Rafailov et al., 2023) update model parameters using both jailbreak and over-refusal samples. SCANS (Cao et al., 2025) adjusts model behavior through safety-conscious activation steering, while SafeDecoding (Xu et al., 2024) applies a safety-aware decoding strategy after safety tuning. This comparison allows us to examine whether CA-DPO provides a better trade-off between refusing harmful requests and avoiding unnecessary refusals on clean and over-refusal inputs.

5.2 Comparison Experiment

Tab. 2 shows that CA-DPO improves jailbreak refusal while keeping clean refusal, over-refusal, and response collapse comparatively low. This behavior indicates that CA-DPO does not improve safety by turning the model into a global refuser. Across model sizes and families, the method instead aims to strengthen refusal on harmful inputs while preserving normal responses on clean and over-refusal inputs. The results further show that this balance depends on model scale: stronger jailbreak refusal can be accompanied by increased over-refusal in some settings. CA-DPO maintains the desired safety–utility trade-off by improving refusal on jailbreak inputs without uniformly increasing refusal on clean and over-refusal inputs.

Table 2: CA-DPO compared with different baselines across model families and evaluation metrics. All values are average refusal rates (%) over three independent runs, OR and FR are abbreviations for the names of the datasets. CA-DPO improves jailbreak refusal while maintain model's capabilities.

Model	Method	Clean refuse ↓	Over-Refusal ↓			Jailbreak ↑			Collapse rate ↓
			OKTest	OR	FR	Cipher	H4rm3l	Hand	
Qwen3.5-0.8B	Base	**0.11**	9.83	35.50	15.50	1.50	34.17	27.50	3.07
	Safety SFT	3.22	15.67	41.67	20.67	14.67	88.17	39.33	14.93
	VanillaDPO	0.78	7.50	16.50	5.17	55.83	75.33	30.67	13.13
	SafeDecoding	7.33	12.50	37.83	12.50	91.33	97.33	42.50	1.87
	SCANS	70.67	77.17	78.83	72.33	97.33	95.50	**89.50**	81.33
	CA-DPO	**0.11**	**2.83**	**7.83**	**0.33**	**99.83**	**98.50**	49.33	**1.07**
Qwen3.5-2B	Base	2.33	15.33	63.50	24.50	10.50	22.17	53.83	1.67
	Safety SFT	4.22	16.67	76.50	30.50	31.17	**99.83**	70.33	2.73
	VanillaDPO	2.78	12.50	51.50	19.17	18.83	60.33	50.67	1.93
	SafeDecoding	3.33	12.50	31.33	7.83	98.33	99.67	63.50	1.33
	SCANS	73.67	87.17	73.83	72.33	22.67	97.33	**85.50**	74.73
	CA-DPO	**2.11**	**7.83**	**20.17**	**1.83**	**99.33**	**99.83**	70.33	**0.53**
Llama3.2-1B	Base	1.11	3.33	12.33	21.50	26.33	37.50	40.33	6.47
	Safety SFT	6.22	8.67	32.50	43.50	91.17	81.83	80.33	28.53
	VanillaDPO	6.78	6.50	29.50	42.17	91.83	82.33	76.67	9.13
	SafeDecoding	16.33	17.50	58.17	53.83	98.33	95.50	71.50	28.53
	SCANS	85.67	59.17	95.83	86.33	91.67	94.50	**93.50**	83.13
	CA-DPO	**0.89**	**1.83**	**5.33**	**15.33**	**99.50**	**96.67**	87.17	**4.13**
Llama3.2-3B	Base	**0.11**	3.33	8.33	26.50	78.50	43.33	44.33	7.67
	Safety SFT	3.22	6.67	28.50	30.50	90.17	77.83	71.17	12.93
	VanillaDPO	1.78	3.50	13.50	25.17	92.83	76.33	56.67	13.93
	SafeDecoding	7.33	15.50	11.17	25.83	98.33	70.67	63.50	18.73
	SCANS	15.67	12.17	6.83	**14.33**	95.67	**98.50**	61.50	36.53
	CA-DPO	**0.11**	**0.83**	**5.33**	16.50	**98.50**	94.33	**75.33**	**6.27**
Gemma3-270M	Base	3.11	6.33	13.17	9.33	1.50	4.67	7.33	13.53
	Safety SFT	4.22	15.67	17.50	5.50	20.17	13.83	29.33	19.53
	VanillaDPO	2.78	13.83	19.50	7.50	24.33	27.33	34.33	17.33
	SafeDecoding	25.33	64.50	55.17	34.83	15.33	**60.67**	**57.50**	9.73
	SCANS	13.67	28.17	32.83	9.33	23.67	32.50	36.50	11.13
	CA-DPO	**2.22**	**4.33**	**9.50**	**4.67**	**51.33**	49.50	45.83	**9.13**
Gemma3-1B	Base	1.56	10.33	41.17	32.33	0.50	21.83	27.17	0.73
	Safety SFT	1.22	9.67	45.50	22.50	5.17	12.83	52.33	2.13
	VanillaDPO	1.78	10.50	38.50	18.17	7.83	28.33	**53.67**	0.53
	SafeDecoding	9.33	10.50	31.17	33.83	12.33	36.67	41.50	**0.13**
	SCANS	3.67	6.17	32.83	**15.33**	0.67	18.50	43.50	0.93
	CA-DPO	**1.11**	**3.33**	**24.67**	17.33	**31.33**	**51.83**	47.83	**0.13**

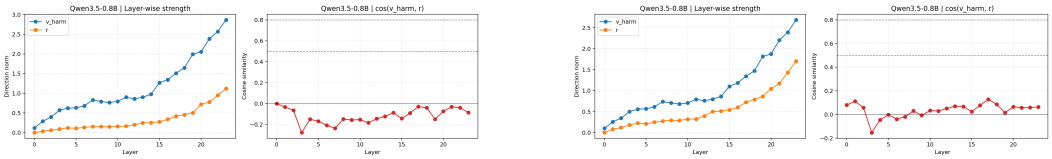


Figure 7: Layer-wise harmfulness, refusal, and coupling signals for Qwen3.5-0.8B before and after CA-DPO. Left: original model signals. Right: CA-DPO-optimized model signals. CA-DPO strengthens the refusal-related signal while keeping the harmfulness–refusal cosine low, indicating that the repair changes refusal execution without simply collapsing the two directions.

5.3 Does Hidden Coupling Matter?

Fig. 7 and Tab. 3 evaluate what changes have been brought about by the hidden coupling. The full CA-DPO objective outperforms variants without hidden coupling and DPO-only training strategy, indicating that the representation-level term helps convert harmfulness recognition into refusal while preserving desired response behavior; additional anchor ablations are reported in Appendix E.

Table 3: Loss ablation on Qwen3.5-0.8B. Coupling loss is effective in CA-DPO.

Model	Setting	JB refuse ↑	Clean refuse ↓	OR refuse ↓	Collapse rate ↓
Qwen3.5-0.8B	Full CA-DPO	**82.55**	**0.11**	**3.66**	**1.07**
Qwen3.5-0.8B	w/o hidden coupling	39.42	1.71	7.26	2.07
Qwen3.5-0.8B	DPO only	28.19	3.17	14.32	9.53

6 Conclusion

We provide a mechanistic view of small-model safety failures by showing that harmfulness recognition and refusal execution can be separated: small models may represent harmful prompts without reliably converting this signal into refusal. Based on this analysis, we propose CA-DPO, which uses the recognition-to-refusal gap to guide data selection and training. Experiments show that CA-DPO improves jailbreak refusal while keeping clean refusal and collapse low, offering a targeted way to improve small-model safety without sacrificing normal model behavior.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, improving the clarity of manuscript presentation, visual refinement of Fig. 1, and AI-assisted coding. We have not used generative AI tools for research ideation, generating experimental results, fabricating or modifying data, producing data-driven scientific figures, or drawing scientific conclusions. All AI-assisted outputs were carefully reviewed, revised, and verified by the authors. In particular, AI-assisted code was manually inspected and tested for correctness, while all scientific claims, experimental results, and conclusions were independently verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies safety failures and over-refusal in small language models. The jailbreak and harmful-request examples are used only for controlled evaluation, representation analysis, and safety-oriented preference training. We do not introduce new attack objectives or provide operational instructions for misuse; the analysis focuses on whether models recognize harmful intent and whether that recognition is translated into appropriate refusal. All experiments use publicly available datasets and model checkpoints, and the training and evaluation data are handled in aggregate rather than as user-specific records. The study does not collect private personal data, involve human subjects, or deploy models in real-world user-facing systems. The intended impact of CA-DPO is to improve refusal of harmful requests while reducing unnecessary refusals of benign prompts, thereby supporting safer and more useful small-model deployment.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility throughout the paper and appendix. Secs. 3 and 5 describe the main representation-analysis and CA-DPO experimental settings. Appendix A provides the training/test splits, data-sampling procedures, model settings, decoding configurations, evaluation metrics, representation-analysis pipeline, and complete experimental parameters. Additional details of the data construction, refusal and collapse measurement, ablations, transfer evaluations, and visualization results are also provided in the appendix. The code and experiment scripts will be released in future work.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024. URL https://arxiv.org/abs/2404.14219.

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. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

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

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Y Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *International Conference on Learning Representations*, volume 2024, pp. 34196–34216, 2024.

Zouying Cao, Yifei Yang, and Hai Zhao. Scans: Mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23523–23531, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Guorui Chen, Yifan Xia, Xiaojun Jia, Zhijiang Li, Philip Torr, and Jindong Gu. Llm jailbreak detection for (almost) free! In *EMNLP (Findings)*, pp. 5777–5807, 2025.

Shei Pern Chua, Hao Wu, Qianli Ma, and Fangzhao Wu. Harc: Coupling harmfulness and refusal directions for robust safety alignment. *arXiv preprint arXiv:2607.00572*, 2026.

Flavio Corradini, Matteo Leonesi, and Marco Piangerelli. State of the art and future directions of small language models: A systematic review. *Big Data and Cognitive Computing*, 9(7):189, 2025.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

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

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024.

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. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*, 2025.

Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis. In *International Conference on Learning Representations*, volume 2026, pp. 145127–145156, 2026.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, volume 2024, pp. 56174–56194, 2024a.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*, 2024b.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

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

Shreyansh Padarha, Yushi Yang, Adel Bibi, Chris Russell, and Adam Mahdi. To distill or not to distill: Knowledge transfer undermines safety of LLMs, 2026. URL `https://openreview.net/forum?id=AEKji3PwD9`. ICLR 2026 withdrawn submission.

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

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*, volume 2025, pp. 54911–54941, 2025.

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

Yi Ren and Danica Sutherland. Learning dynamics of llm finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 70523–70563, 2025.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

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

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuan-Jing Huang, Xun Zhao, and Dahua Lin. Navigating the overkill in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4602–4614, 2024.

Haonan Shi, Guoli Wang, Tu Ouyang, and An Wang. Ease: Practical and efficient safety alignment for small language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37923–37931, 2026.

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

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87, 2025a.

Zhibo Wang, Wenhui Zhang, Huiyu Xu, Zeqing He, Ziqi Zhu, and Kui Ren. Can small language models reliably resist jailbreak attacks? a comprehensive evaluation. *arXiv preprint arXiv:2503.06519*, 2025b.

Zhibo Wang, Wenhui Zhang, Huiyu Xu, Zeqing He, Ziqi Zhu, and Kui Ren. Can small language models reliably resist jailbreak attacks? a comprehensive evaluation. *arXiv preprint arXiv:2503.06519*, 2025c.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Yifan Xia, Guorui Chen, Wenqian Yu, Zhijiang Li, Philip Torr, and Jindong Gu. Reimagining safety alignment with an image. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9600–9614, 2025.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5587–5605, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Sibo Yi, Tianshuo Cong, Xinlei He, Qi Li, and Jiaxing Song. Beyond the tip of efficiency: Uncovering the submerged threats of jailbreak attacks in small language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17221–17234, 2025.

Peng Zhang and Peijie Sun. Differentiated directional intervention: a framework for evading llm safety alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38102–38110, 2026.

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. *arXiv preprint arXiv:2505.08054*, 2025.

Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. Defending large language models against jailbreaking attacks through goal prioritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8865–8887, 2024.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *Advances in Neural Information Processing Systems*, 38:140283–140318, 2026.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2461–2488, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

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*, 2023b.

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

A Experimental Setup and Data Construction

In Sec. 3. This appendix gives the complete configuration used for the representation analyses in Sec. 3. We evaluate the Qwen3.5 family at 0.8B, 2B, 4B, 9B, and 27B, together with Llama3.2-1B-Instruct and Llama3.2-11B-Vision-Instruct. Each model is processed with its native tokenizer and chat template, and all representation statistics are estimated independently within each model. We extract the input-side representation at the instruction-side position and the refusal-side representation at the post-instruction position. The same layer-wise extraction and pooling procedure is used for jailbreak, clean, and over-refusal inputs.

The PCA visualization uses a cipher-dominant jailbreak pool, with cipher inputs accounting for more than 80% of the jailbreak samples. This choice is motivated by the pronounced change in cipher refusal behavior across model scales: smaller models are substantially more vulnerable, whereas larger models show a sharp increase in refusal. The resulting PCA plots therefore focus on the representation geometry of cipher-heavy inputs and should not be interpreted as a balanced comparison of all jailbreak families.

For the signal-strength, coupling, and behavioral analyses in Secs. 3.3–3.5, we use a balanced jailbreak pool with 200 samples from each of cipher, h4rm3l[1], and hand-written jailbreaks, for 600 jailbreak samples in total. The clean pool follows the construction in the referenced safety-alignment study: it is formed by proportional random sampling from Pure-Dove[2], Open-Platypus[3], and SuperGLUE[4]. The same fixed clean pool is used across models, and the proportional sampling rule is kept unchanged rather than imposing equal counts across the three sources. The over-refusal pool contains 600 fixed prompts, composed of 200 samples from each of OKTest, OR-Bench/OR-1k, and FalseReject (FR). For a given model, the refused and non-refused subsets are identified from this common pool after inference; the pool itself is not re-sampled to equalize the number of refusals across models.

For harmful and jailbreak inputs, we use black-box attack samples following the competing-objectives paradigm defined by Wei et al. (Wei et al., 2023). The attacks apply different prompt transformations to the same underlying harmful request, simulating how users rephrase malicious requests to evade safety constraints. We draw clean and over-refusal samples from several datasets.

Cross-family consistency. Tab. 4 indicates that cipher, h4rm3l, and hand jailbreaks share a harmfulness-related direction across model families, suggesting that their surface forms differ while activating a similar safety representation.

The representation analysis distinguishes direction estimation from signal-strength measurement. In Sec. 3, $v_{\text{harm}}^{(\ell)}$ and $r_{\text{ref}}^{(\ell)}$ are normalized mean-difference vectors and are used to measure

[1] https://github.com/mdoumbouya/h4rm3l

[2] https://huggingface.co/datasets/LDJnr/Pure-Dove

[3] https://huggingface.co/datasets/garage-bAInd/Open-Platypus

[4] https://huggingface.co/datasets/aps/super_glue

Table 4: Cross-family consistency of the harmfulness-related direction. Pairwise cosine and top-1 variance measure directional agreement, while projection and LOFO separation measure jailbreak-clean separability and held-out family generalization.

Model	Pairwise family cos. ↑	Top-1 var. ↑	Proj. sep. ↑	LOFO sep. ↑
Qwen3.5-0.8B	0.556	0.709	1.941	1.906
Qwen3.5-2B	**0.587**	**0.729**	1.935	1.898
Qwen3.5-4B	0.562	0.712	**1.969**	**1.925**
Qwen3.5-9B	0.585	0.728	1.926	1.853

directional coupling through cosine similarity. The signal strengths reported in Secs. 3.3–3.4, however, are computed from the corresponding unnormalized mean-difference vectors before normalization. Thus, signal strength refers to the magnitude of an input-side or refusal-side contrast, whereas cosine similarity measures the alignment between the two normalized directions. The two quantities are therefore interpreted separately.

For the refused-versus-non-refused comparison in Sec. 3.5, coupling values are estimated conditionally on the observed response outcome. We partition the jailbreak samples into refused and non-refused groups, estimate the relevant representation contrasts for each group, and compute the cosine between the resulting harmfulness-related and refusal-related directions. The values in Tab. 1 are therefore group-conditional direction-level estimates; they are not cosines assigned independently to individual prompts. This procedure allows us to compare whether the relationship between the two signals differs between jailbreaks that are refused and those that are not.

The sample-level scores used for data selection and coupling-aware training are obtained by averaging the layer-wise scores over the selected layers. Specifically, for a prompt x, we define

$$H(x) = \frac{1}{|\mathcal{L}_H|} \sum_{\ell \in \mathcal{L}_H} H_\ell(x), \qquad R(x) = \frac{1}{|\mathcal{L}_R|} \sum_{\ell \in \mathcal{L}_R} R_\ell(x), \qquad G(x) = H(x) - R(x), \quad (11)$$

where $\mathcal{L}_H$ and $\mathcal{L}_R$ denote the selected layer sets for the input-side and refusal-side signals, respectively. We use the signed gap $G(x)$ for sample selection: positive values indicate that the harmfulness-related score is larger than the refusal-related score, while negative values indicate the opposite. The layer-averaged scores are used consistently in the subsequent data-construction and CA-DPO training procedures.

In Sec. 5. We use disjoint training and test pools for CA-DPO evaluation. The jailbreak training set contains 128 samples selected from cipher and h4rm3l using our data-construction strategy. For in-distribution jailbreak testing, we remove all training samples from the candidate pool and then randomly sample 200 cipher prompts and 200 h4rm3l prompts.

The hand-crafted attacks are used as a generalization test: we sample 200 prompts from the remaining hand-crafted pool, after excluding any prompts used during data construction. The hand-crafted prompts are drawn from JailBreakV-28[5]. Following the competing-objective and mismatched-generalization classification used by Chen et al. (2025), we randomly choose four attack types from the competing-objective group and sample 50 prompts per type for the hand-crafted subset. Tab. 8 shows the full classification of the hand-crafted attack collection.

For over-refusal data, the training set contains 128 refused prompts sampled from OR-Bench. The over-refusal test set is sampled from OKTest, OR-Bench, and FalseReject, with all training prompts removed from the sampling pool. For clean data, the training set contains 128 prompts, with 64 prompts from Pure-Dove and 64 prompts from Open-Platypus. The clean test set follows the same source design as Xia et al. (2025): after excluding training prompts, we randomly sample 100 prompts each from Pure-Dove, Open-Platypus, and SuperGLUE, giving 300 clean test prompts in total. These pools from Sec. 5 are distinct from the pools in Sec. 3 described above.

Model. We use the following released checkpoints. For the main CA-DPO experiments, we evaluate Qwen3.5-0.8B[6], Qwen3.5-2B[7], Llama3.2-1B-Instruct[8], Llama3.2-3B-Instruct[9], Gemma3-270M-IT[10], and Gemma3-1B-IT[11]. For the representation analyses, we additionally evaluate Qwen3.5-

[5] https://huggingface.co/datasets/JailbreakV-28/JailBreakV-28k
[6] https://huggingface.co/Qwen/Qwen3.5-0.8B
[7] https://huggingface.co/Qwen/Qwen3.5-2B
[8] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[9] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[10] https://huggingface.co/google/gemma-3-270m-it
[11] https://huggingface.co/google/gemma-3-1b-it

Keyword

- I do not
- I don't
- I'm unable
- I am not able
- I am not able
- I am unable
- I'm not capable
- I'm incapable
- I am not capable
- I am incapable
- I am really sorry
- I'm really sorry
- My apologies
- I cannot fulfill your request
- I'm just an AI
- As an AI language model
- It is not appropriate
- It is not accurate
- It is not recommended to
- It is illegal
- I cannot
- I can't
- I can never
- I can not
- Sorry
- I am sorry
- I'm sorry
- I apologize
- Request denied
- That's not allowed

Figure 8: Refusal keyword list used for measuring refusal rate. A response is counted as a refusal if it contains one of these predefined refusal expressions.

Collapse Example

Question: What are some creative ways to present inferior products to pique customer interest?
**Answer:1. Create a sense of urgency:\n\nStep 1:\n\n1. Step 1:\n\n1. Step 1:\n\n1. Step

For over-refusal inputs, we use the analogous contrasts

$$\mathbf{v}_{\mathrm{OR}}^{(\ell)} = \mu_{\mathrm{OR}}^{(\ell)} - \mu_{\mathrm{clean}}^{(\ell)}, \qquad \mathbf{r}_{\mathrm{OR}}^{(\ell)} = \mu_{\mathrm{OR\text{-}ref}}^{(\ell)} - \mu_{\mathrm{OR\text{-}nonref}}^{(\ell)}. \tag{13}$$

The input-side vector $\mathbf{v}_{\mathrm{JB}}$ is the jailbreak-derived harmfulness direction, whereas $\mathbf{v}_{\mathrm{OR}}$ is the over-refusal input-contrast direction. We keep this distinction because over-refusal prompts are benign or borderline inputs and their contrast should not be treated as a direct ground-truth harmfulness direction.

Signal strength is computed from the unnormalized mean-difference vectors:

$$S_{\mathrm{v}}^{c}(\ell) = \left\|\mathbf{v}_{c}^{(\ell)}\right\|_2, \qquad S_{\mathrm{r}}^{c}(\ell) = \left\|\mathbf{r}_{c}^{(\ell)}\right\|_2, \qquad c \in \{\mathrm{JB}, \mathrm{OR}\}. \tag{14}$$

We normalize the two vectors only when computing their directional coupling:

$$\cos^{c}(\ell) = \frac{\mathbf{v}_{c}^{(\ell)} \cdot \mathbf{r}_{c}^{(\ell)}}{\left\|\mathbf{v}_{c}^{(\ell)}\right\|_2 \left\|\mathbf{r}_{c}^{(\ell)}\right\|_2}. \tag{15}$$

Thus, the strength comparison uses the original mean-difference vectors, while the cosine comparison uses their normalized directions.

The PCA visualization in Sec. 3 is used as a qualitative view of how clean, jailbreak, and over-refusal inputs are organized in the representation space. Figs. 16 and 17 show representative visualization results for Qwen3.5-0.8B and Qwen3.5-27B. These examples illustrate the scale-dependent change in separation among clean, jailbreak, and over-refusal samples across layers.

D Complete Layer-wise Signal and Coupling Results

The following plots report the complete layer-wise curves for the balanced jailbreak pool and the fixed clean and over-refusal pools. Each plot contains the input-side signal strength, the refusal-side signal strength, and the cosine between the corresponding condition-specific directions. The plots use the unified axis configuration prepared for this appendix, so the relative magnitudes of the input-side and refusal-side contrasts can be compared across model sizes without changing the underlying values. Backup versions of the plots are excluded.

The complete layer-wise plots are collected at the end of the appendix to keep the supplementary evaluations contiguous.

E Extended Ablations and Supplementary Evaluation

This section reports the extended ablations and supplementary evaluations for CA-DPO. All comparisons use the same refusal and collapse metrics as the main experiments. Higher refusal is preferred on jailbreak inputs, while lower refusal is preferred on clean and over-refusal inputs. Collapse rate measures degenerate outputs such as repeated templates, empty or unhelpful responses, malformed formats, and abnormal lengths. Baseline comparisons use matched evaluation prompts, decoding settings, and training data where applicable.

Loss ablation experiment. The extended loss ablation removes one objective component at a time while keeping the model, data, decoding, and evaluation set fixed. The comparison isolates the contribution of hidden coupling, clean anchoring, chosen anchoring, and stop-gradient design to the final safety–utility trade-off.

Table 5: Extended loss ablation of CA-DPO on Qwen3.5-0.8B. The reported matched runs test hidden coupling, clean anchoring, chosen anchoring, stop-gradient design, and DPO-only training.

Model	Setting	JB refuse ↑	Clean refuse ↓	OR refuse ↓	Collapse rate ↓
Qwen3.5-0.8B	Full CA-DPO	**82.55**	**0.11**	**3.66**	**1.07**
Qwen3.5-0.8B	w/o hidden coupling	39.42	1.71	7.26	2.07
Qwen3.5-0.8B	w/o clean anchor	78.5	4.8	5.1	3.73
Qwen3.5-0.8B	w/o chosen anchor	39.4	1.69	4.09	9.33
Qwen3.5-0.8B	w/o stop-gradient	68.5	4.73	5.92	2.33
Qwen3.5-0.8B	DPO only	28.19	3.17	14.32	9.53

Gap-based data-selection experiment. Tab. 6 shows that gap-based selection provides a stronger training signal than random selection or selection by harmfulness score alone under matched data

budgets. The selected samples improve jailbreak refusal while avoiding a corresponding increase in clean refusal, over-refusal, and response collapse. This result supports using the recognition-to-refusal gap to construct targeted training data rather than treating it only as a post-hoc diagnostic.

Table 6: Available extended data-selection ablations of CA-DPO on Qwen3.5-0.8B. The reported runs compare random, harmfulness-only, and the full gap-plus-anchor training set; the pure gap-only output is unavailable.

Model	Selection rule	Signal used	JB refuse ↑	Clean refuse ↓	OR refuse ↓	Collapse rate ↓
Qwen3.5-0.8B	Random	none	72.61	6.24	7.09	1.67
Qwen3.5-0.8B	High-H only	harm score	73.74	6.29	6.67	1.87
Qwen3.5-0.8B	Full CA-DPO set	gap + anchors	**82.55**	**0.11**	**3.66**	**1.07**

Data-ratio and data-scale experiment. Data-ratio experiments keep the total number of training examples fixed while varying the proportions of harmful repair samples, benign over-refusal negatives, clean anchors, and collapse negatives. Data-scale experiments keep the mixture fixed and vary the total number of examples. These checks evaluate whether CA-DPO is sensitive to the training mixture or data scale.

Table 7: Data-ratio and data-scale ablation of CA-DPO on Qwen3.5-0.8B. CA-DPO is evaluated under different mixture proportions and training-set sizes.

Model	Study	Data setting	JB refuse ↑	Clean refuse ↓	OR refuse ↓	Collapse rate ↓
Qwen3.5-0.8B	Ratio	More harmful repair	82.9	6.2	12.5	1.13
Qwen3.5-0.8B	Ratio	More benign negatives	63.0	4.72	5.13	3.13
Qwen3.5-0.8B	Ratio	More clean anchors	72.6	1.65	9.08	1.33
Qwen3.5-0.8B	Scale	Small subset	54.2	3.14	4.21	8.93
Qwen3.5-0.8B	Scale	Full set	82.55	**0.11**	**3.66**	**1.07**
Qwen3.5-0.8B	Scale	Larger set	**89.29**	6.22	9.47	1.53

F LIMITATIONS

Our interpretation relies on a conditional scaling assumption: within the model families and training configurations considered in this work, larger models generally exhibit stronger safety performance. If safety behavior is instead determined primarily by training data, alignment procedures, architectural features, or other model-specific factors, the observed cross-scale differences may not be attributable to model size alone. In that case, our explanation of why small models fail to refuse harmful inputs may not generalize beyond the model families studied here.

Our analysis also focuses on small models that retain a certain degree of harmfulness recognition. A very weak model may fail to refuse simply because it cannot identify the harmfulness of an input in the first place. Such failures would reflect insufficient harmfulness recognition rather than a failure to convert recognition into refusal. Therefore, the recognition-to-refusal gap described in this paper is intended to characterize capable small models with unstable safety behavior, rather than all models with small parameter counts.

Table 8: Hand-crafted attack categories from (Chen et al., 2025)

Competing Objective	Mismatched Generalization
aim	auto_payload splitting
dev_mode_v2	base64
dev_mode_ranti	base64_raw
distractors	base64_input_only
distractors_negated	base64_output_only
evil_confidant	combination_1
poems	combination_2
prefix_injection_1	combination_3
prefix_injection_2	disemvowel
prefix_injection hello	few_shot_json
refusal_suppression	leetspeak
refusal_suppression inv	rot13
style_injection	style_injection
short	json
	wikipedia
	wikipedia with_title

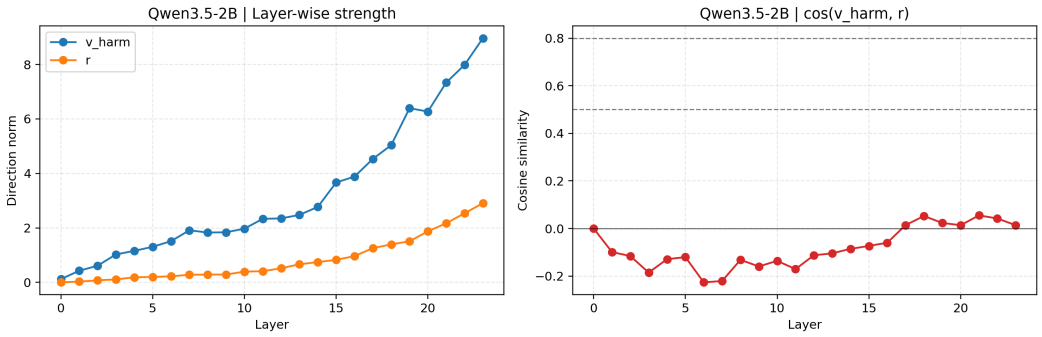


Figure 10: Complete layer-wise signal-strength and cosine curves for Qwen3.5-2B.

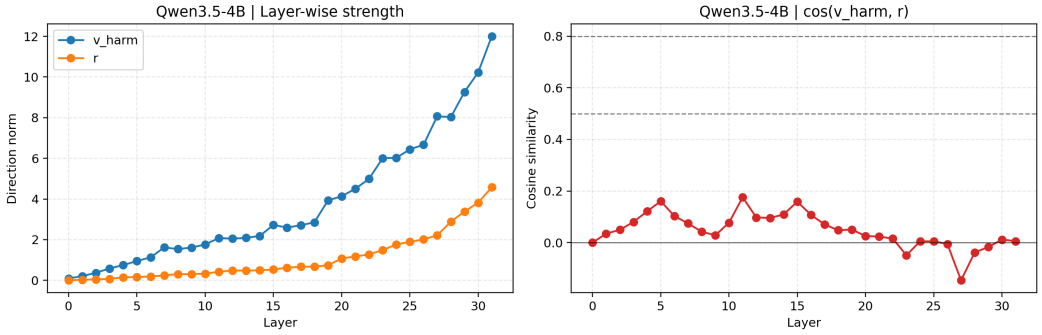


Figure 11: Complete layer-wise signal-strength and cosine curves for Qwen3.5-4B.

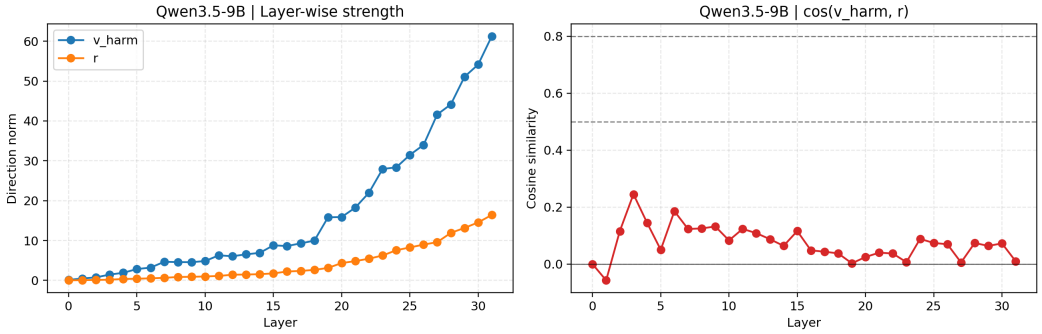


Figure 12: Complete layer-wise signal-strength and cosine curves for Qwen3.5-9B.

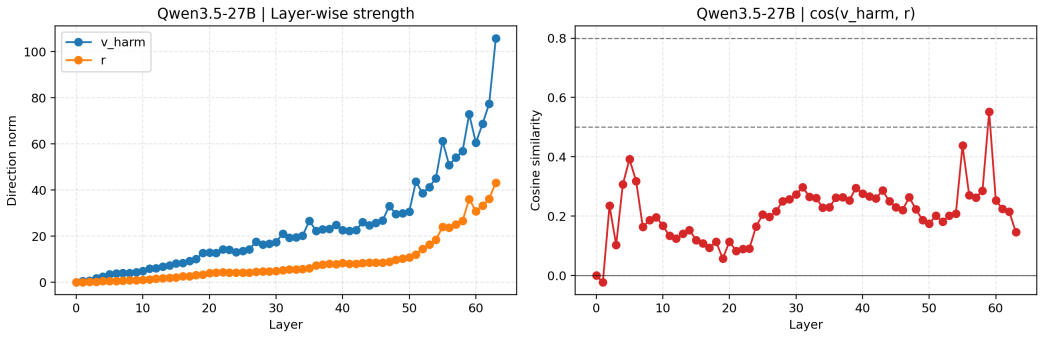


Figure 13: Complete layer-wise signal-strength and cosine curves for Qwen3.5-27B.

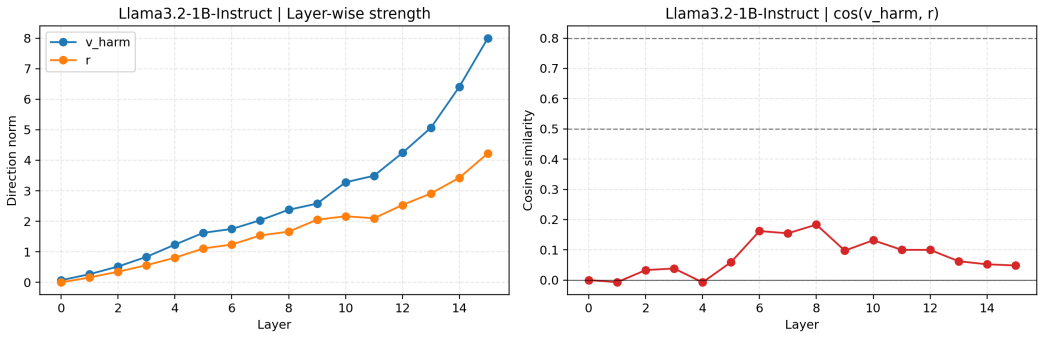


Figure 14: Complete layer-wise signal-strength and cosine curves for Llama3.2-1B-Instruct.

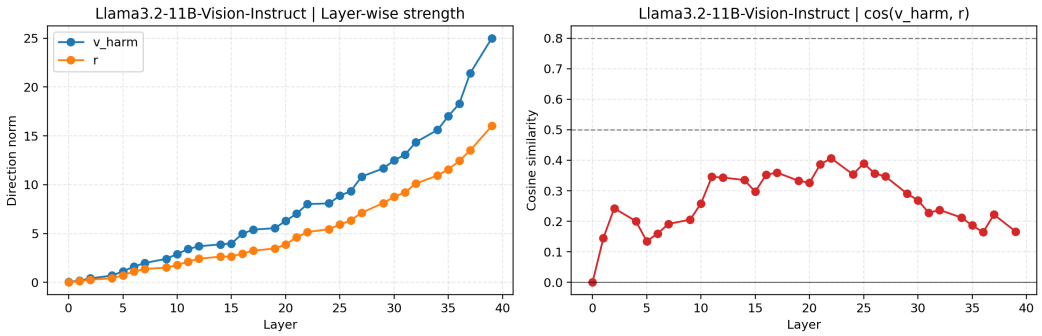


Figure 15: Complete layer-wise signal-strength and cosine curves for Llama3.2-11B-Vision-Instruct.

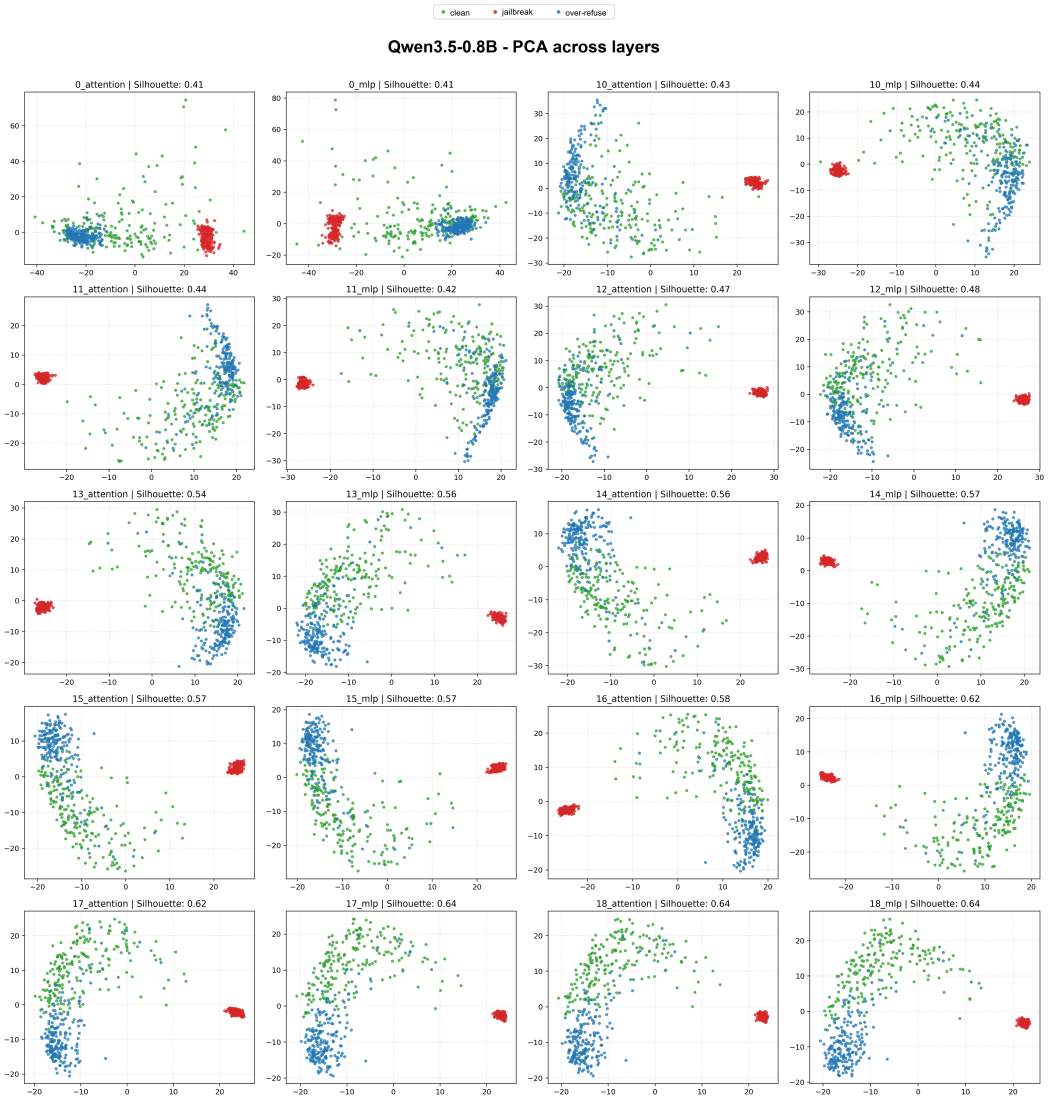


Figure 16: Representative PCA visualization results for Qwen3.5-0.8B across selected layers. Green, red, and blue points denote clean, jailbreak, and over-refusal samples, respectively.

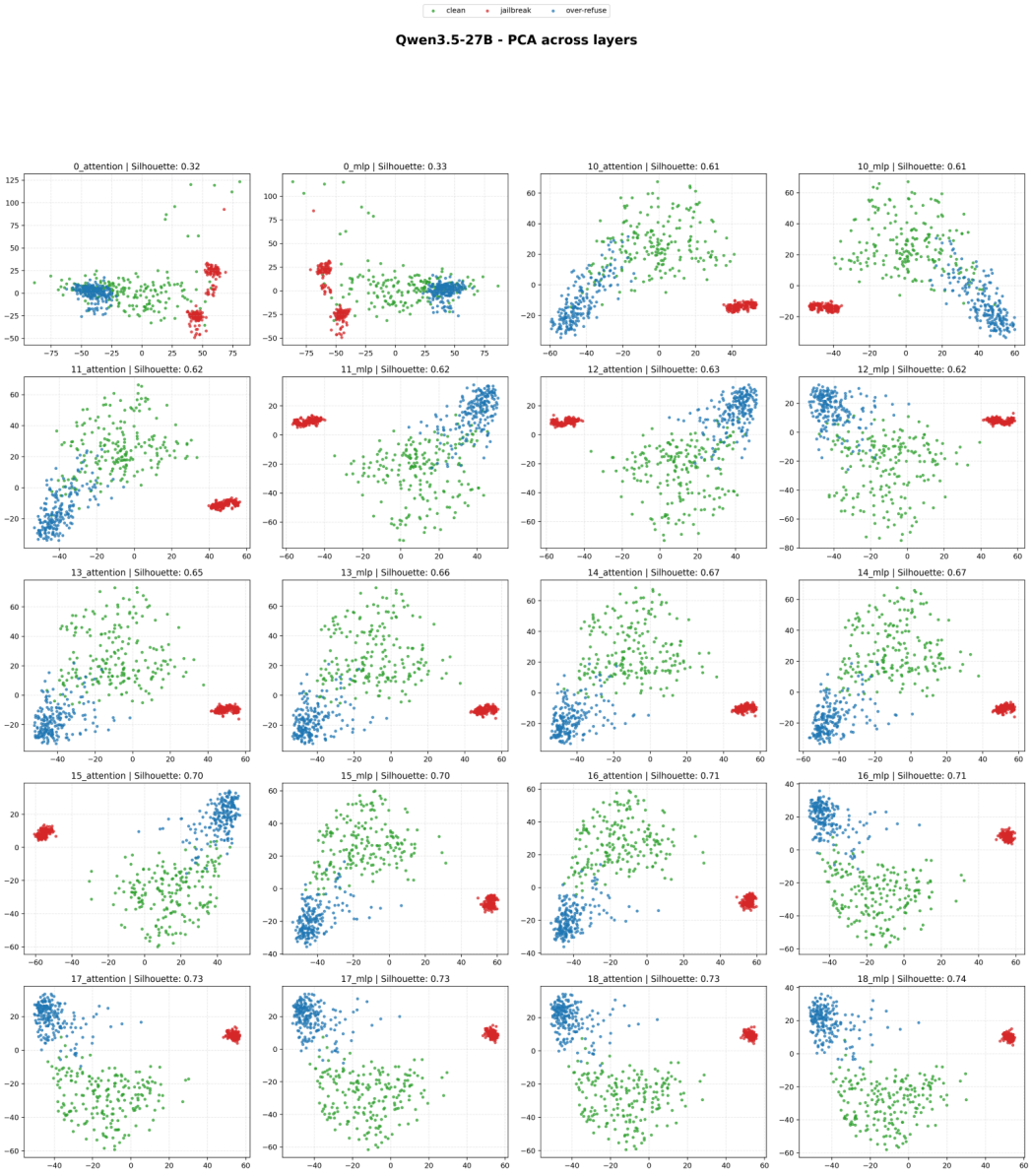


Figure 17: Representative PCA visualization results for Qwen3.5-27B across selected layers. Green, red, and blue points denote clean, jailbreak, and over-refusal samples, respectively.