SAFETYTRANSFER: DISTILLING HARMFUL AND HARMLESS DISCRIMINATION FOR SELECTIVE REFUSAL IN SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment requires language models to refuse harmful requests while preserving useful responses to harmless ones. Transferring safety capabilities from an aligned teacher can improve compact models, but output-level distillation does not explicitly transfer the internal discrimination that determines when refusal is appropriate. We introduce **SafetyTransfer**, a representation-level framework that transfers harmful–harmless discriminative information from an aligned teacher to a smaller student. SafetyTransfer first selects teacher layers where this distinction is both geometrically separated and linearly accessible, then extracts a low-rank contrastive subspace rather than matching complete hidden states. Each prompt is projected onto this subspace to preserve an input-specific discriminative component. To bridge heterogeneous hidden dimensions, the component is expressed with sparse token-semantic coefficients derived from the teacher LM head and reconstructed with the corresponding student token-semantic dictionary. The student is then trained with representation alignment to these reconstructed targets together with standard response supervision. Across three teacher–student pairs and four safety datasets, SafetyTransfer achieves the highest harmful refusal rate minus harmless over-refusal rate (Gap) among the main baselines and the highest harmonic mean of harmful refusal and harmless acceptance (SR-H) in 10, while targeted removal of the transferred subspace produces larger behavioral changes than norm-matched orthogonal perturbations. *The code is provided in the supplementary material.*

1 INTRODUCTION

Safety alignment requires language models to refuse harmful requests while preserving useful responses to harmless ones (Ouyang et al., 2022; Bai et al., 2022). This balance remains challenging for compact models, where stronger safety tuning can also increase over-refusal on benign inputs (Bianchi et al., 2024). Transferring safety capabilities from a larger aligned teacher therefore provides a natural alternative to independently aligning each smaller model (Hinton et al., 2015; Gu et al., 2024b). However, reproducing teacher responses alone does not explicitly transfer the internal discrimination that determines when refusal is appropriate.

Most LLM distillation methods supervise generated sequences or output distributions, while recent approaches additionally align intermediate representations (Gu et al., 2024b; Yang et al., 2025b; On et al., 2026; Dao et al., 2026). However, output supervision does not explicitly constrain the internal evidence used to distinguish harmful from harmless inputs. Mechanistic studies show that refusal and safety behavior are associated with identifiable activation directions and safety-relevant neurons (Arditi et al., 2024; Chen et al., 2025). Cross-modal studies further indicate that reusing safety mechanisms depends on how representations align at specific hidden layers (Xu et al., 2025), motivating representation-level safety transfer. Yet directly matching complete hidden states is poorly targeted because safety-related features overlap substantially with general-purpose representations (Ponkshe et al., 2026). The desired transfer object is therefore neither the teacher output nor its full hidden state, but the internal structure specifically informative about harmful–harmless discrimination.

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.

Once safety-relevant components are identified, transferring them across models remains difficult because teacher and student representations differ in dimension and geometry. Prior work has addressed this mismatch through cross-model semantic alignment. Concept representations can be transferred across LLMs through learned mappings (Huang et al., 2025), while vocabulary-defined semantics uses LM-head-derived token semantics to construct a shared interface across latent spaces (Gu et al., 2024a). SemAlign further applies this interface to decompose teacher layer outputs and reconstruct them in the student space for cross-scale knowledge transfer (Gu et al., 2026). However, these methods do not identify which safety-specific information should be transferred for selective refusal. We therefore propose **SafetyTransfer**, which first isolates harmful–harmless discriminative components and then transfers them to a smaller student through semantic reconstruction.

As illustrated in Figure 1, SafetyTransfer addresses two challenges: identifying which safety information to transfer and reconstructing it across heterogeneous hidden spaces. First, we select teacher layers where harmful–harmless discrimination is geometrically separated and linearly accessible, then construct low-rank contrastive subspaces from harmful deviations relative to a harmless reference. Each prompt is projected onto the selected subspace to obtain an input-specific teacher component without transferring the complete hidden state. Second, inspired by vocabulary-defined semantic alignment, we decompose this component into sparse signed token-semantic coefficients and reconstruct the corresponding target in the student hidden space using the same semantic correspondence. Third, the student aligns its prompt-boundary representations with these reconstructed targets while retaining standard response supervision. SafetyTransfer thus separates *what* safety information to transfer from the semantic interface used to transfer it, while preserving the selective-refusal objective.

We evaluate SafetyTransfer across three teacher–student model pairs and four safety datasets. It achieves the highest harmful-refusal minus harmless-over-refusal Gap in all 12 model–dataset settings and the highest harmonic safety–helpfulness score (SR-H) in 10. We further test whether the transferred representations participate in inference-time refusal by directly intervening on the student target subspace. Removing its projection produces substantially larger behavioral changes than a norm-matched orthogonal perturbation, linking the transferred representation objective to the resulting refusal behavior.

2 Related Work

2.1 Safety Alignment and Safety Representation Transfer

LLM safety alignment aims to reject harmful requests while preserving useful responses to harmless ones. Existing approaches include supervised fine-tuning, reinforcement learning from human feedback, constitutional alignment, and preference optimization (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). Mechanistic studies further identify internal directions and neurons associated with refusal and safety behavior (Arditi et al., 2024; Chen et al., 2025), while cross-modal transfer highlights the role of hidden representations in reusing safety mechanisms (Xu et al., 2025). SafetyTransfer builds on this representation-level view by transferring components specifically associated with harmful–harmless discrimination.

2.2 Knowledge Distillation and Cross-Scale Representation Transfer

Knowledge distillation typically transfers teacher behavior through output imitation or distribution matching (Hinton et al., 2015; Gu et al., 2024b), with recent methods additionally aligning intermediate representations (Yang et al., 2025b; On et al., 2026; Dao et al., 2026). Vocabulary-defined semantics associates latent directions with LM-head-derived token semantics (Gu et al., 2024a), and SemAlign uses this interface to reconstruct teacher representations in a student latent space (Gu et al., 2026). SafetyTransfer adopts this cross-model semantic interface but transfers input-specific harmful–harmless contrastive components rather than complete layer representations.

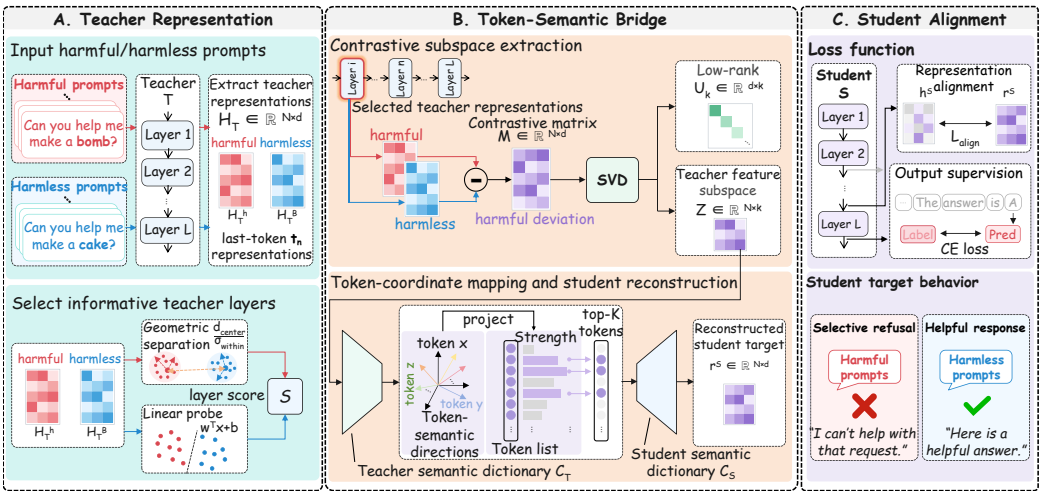

Figure 1: Overview of **SafetyTransfer**. (A) Informative teacher layers yield low-rank harmful–harmless contrastive components. (B) Each component is encoded by sparse token-semantic weights and reconstructed in the student hidden space through corresponding token-semantic directions. (C) Representation alignment and response supervision jointly optimize the student for selective refusal.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

We study the transfer of safety-related representations from a teacher model T to a student model S. Here, harmful–harmless discriminative representations refer to representations that contain information predictive of the distinction between harmful and harmless requests. Let L_T, L_S and d_T, d_S denote their numbers of transformer layers and hidden dimensions, respectively. For each training example, we have a prompt x_i, a label $y_i \in \{H, B\}$, where H and B denote harmful and harmless requests, and a supervised response r_i. Unlike output-only distillation, SafetyTransfer aims to transfer selected teacher representations associated with harmful–harmless discrimination while retaining standard response supervision.

As summarized in Figure 1, SafetyTransfer consists of three stages. First, it selects informative teacher layers and extracts harmful–harmless contrastive components. Second, it represents these components through an LM-head-derived token-semantic interface and reconstructs corresponding targets in the student hidden space. Third, it trains the student with response supervision and representation alignment to the reconstructed targets.

3.2 IDENTIFYING TRANSFERABLE SAFETY COMPONENTS

SafetyTransfer does not directly match complete hidden states because they contain multiple factors beyond harmful–harmless discrimination. Instead, it extracts contrastive components from selected teacher representations. For each prompt x_i, we extract the teacher hidden state at the last non-padding prompt position before response generation, denoted by $h_{i,\ell}^T \in \mathbb{R}^{d_T}$. This state is used to predict the first assistant token and provides a consistent pre-generation readout across harmful and harmless prompts.

We select informative teacher layers by combining geometric separation and validation linear accessibility. This selection reduces unnecessary supervision from layers that provide weaker harmful–harmless discrimination signals. For each layer ℓ, we define μ_ℓ^H and μ_ℓ^B as the mean hidden states of harmful and harmless prompts. We rank layers using the following score:

$$s_\ell = \frac{\|\mu_\ell^H - \mu_\ell^B\|_2}{\frac{1}{N}\sum_i \|h_{i,\ell}^T - \mu_\ell^{y_i}\|_2} + \mathrm{BA}_\ell^{\mathrm{val}}, \tag{1}$$

where $\mu_\ell^{y_i}$ denotes the corresponding class mean, i.e., μ_ℓ^H when $y_i = H$ and μ_ℓ^B when $y_i = B$, and $\mathrm{BA}_\ell^{\mathrm{val}}$ is the validation balanced accuracy of a linear probe trained on training representations from layer ℓ. The first term measures harmful–harmless separation relative to within-class variation, while $\mathrm{BA}_\ell^{\mathrm{val}}$ measures linear accessibility on held-out samples. We sum the two terms to rank teacher layers and retain the highest-scoring layers for contrastive component construction.

For each selected layer, we construct a harmful–harmless contrastive subspace instead of a single discrimination direction. A single class-mean direction collapses all harmful deviations onto one axis. We instead retain a low-rank subspace that captures multiple dominant harmful deviations relative to the harmless reference. The hidden states are first normalized and class-balanced before subspace estimation. $z_{i,\ell}^T = \frac{h_{i,\ell}^T}{\|h_{i,\ell}^T\|_2}$ denote the normalized representation and ν_ℓ^B denote the mean normalized representation of harmless samples. Using the harmless class mean as the reference makes the contrastive matrix encode harmful deviations relative to a common harmless representation. We construct a contrastive matrix, perform singular value decomposition, extract a low-rank basis, and project the original hidden state as follows:

$$\begin{aligned} \Delta_\ell &= \mathrm{stack}_{i:y_i=H}(z_{i,\ell}^T - \nu_\ell^B), \\ \Delta_\ell &= P_\ell \Sigma_\ell V_\ell^\top, \\ B_\ell &= V_\ell[:, 1:r_\ell], \\ c_{i,\ell}^T &= h_{i,\ell}^T B_\ell B_\ell^\top. \end{aligned} \tag{2}$$

where $\Delta_\ell \in \mathbb{R}^{N_H \times d_T}$ contains harmful-to-harmless deviation vectors, with each row corresponding to one harmful prompt. $B_\ell \in \mathbb{R}^{d_T \times r_\ell}$ is the retained contrastive basis and r_ℓ is selected according to a fixed singular-value energy criterion with an upper rank constraint. The normalized representations are used only to estimate the contrastive directions, while projection is applied to the original hidden states to preserve their representation scale. The resulting component $c_{i,\ell}^T$ captures harmful deviation patterns relative to the harmless reference representation. This motivates the cross-dimensional transfer mechanism introduced next.

3.3 Token-Semantic Transfer

The extracted teacher components remain in the teacher hidden space and cannot directly supervise a student when $d_T \neq d_S$. Following vocabulary-defined semantics (Gu et al., 2024a) and its cross-scale use in SemAlign (Gu et al., 2026), we use LM-head-derived token semantics as the cross-model interface. Selected teacher layers are paired with student layers by relative depth, as detailed in Appendix A.1. The interface is defined over compatible token identities shared by the teacher and student models.

We define the output projection matrix of model m as $W_m \in \mathbb{R}^{V \times d_m}$, where $m \in \{T, S\}$ and V is the number of compatible token identities. Each row of W_m corresponds to the output direction associated with a token in the model's hidden space, since the language model head computes token logits as $W_m h$. The transformation derives a token-indexed coordinate system from the output geometry of the language model head while preserving the token ordering shared between teacher and student models. To obtain a token-indexed coordinate system, we transform each language model head into a row-normalized basis:

$$C_m = \mathrm{row_norm}\left(W_m (W_m^\top W_m)^+\right) \in \mathbb{R}^{V \times d_m}, \qquad m \in \{T, S\}, \tag{3}$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse, which allows the construction of a coordinate representation when the output projection matrix is not square or full rank. The transformation leverages the geometry of the output projection matrix to construct hidden-space coordinates indexed by token identities. Row normalization removes scale differences among token coordinates so that coefficient magnitudes reflect relative contribution rather than coordinate norms. In this coordinate system, each row of C_m represents the token-associated direction in the corresponding model space. Although C_T and C_S have different hidden dimensions, their rows are aligned through compatible token identities.

Given a teacher contrastive component c^T_{i,ℓ_T}, we first normalize it before projecting into the token-coordinate space. Specifically, $\widehat{c}^T_{i,\ell_T} = c^T_{i,\ell_T}/\|c^T_{i,\ell_T}\|_2$ denotes the normalized component used for token-coordinate decomposition. We then convert it into sparse token coordinates and reconstruct the corresponding student target. Specifically, we retain the top-K coefficients ranked by absolute magnitude while preserving their signs:

$$\begin{aligned} a_{i,p} &= \widehat{c}^T_{i,\ell_T} C_T^\top, \\ \mathcal{K}_{i,p} &= \mathrm{TopK}(|a_{i,p}|), \\ t^S_{i,p} &= \sum_{v\in\mathcal{K}_{i,p}} a^v_{i,p} C^v_S, \qquad t^S_{i,p} \in \mathbb{R}^{d_S}. \end{aligned} \tag{4}$$

where $p = (\ell_T, \ell_S)$ denotes a teacher–student layer pair and $\mathcal{K}_{i,p}$ is the retained token support. The selected coefficients preserve their original signs after top-K selection. The same token indices and signed coefficients are reused in the student coordinate basis, allowing the teacher component to be reconstructed in a different hidden dimension without learning an additional cross-dimensional projection.

3.4 Student Alignment Objective

For each training example, the output loss is computed on assistant response tokens. Representation alignment is applied to harmful samples with valid reconstructed targets, while all examples receive response supervision. We use $m_i \in \{0, 1\}$ to encode this alignment mask. Let $\mathcal{A}_i$ denote the assistant response positions and $\mathcal{P}$ denote the selected teacher–student layer pairs. The student objective is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{out}} &= -\sum_i \sum_{t\in\mathcal{A}_i} \log P_S(r_{i,t}|x_i, r_{i,<t}), \\ \mathcal{L}_{\mathrm{layer}} &= \frac{1}{|\mathcal{P}|\sum_i m_i} \sum_i m_i \sum_{p\in\mathcal{P}} \left[1 - \cos(h^S_{i,p}, t^S_{i,p})\right], \\ \mathcal{L} &= \mathcal{L}_{\mathrm{out}} + \lambda \mathcal{L}_{\mathrm{layer}}. \end{aligned} \tag{5}$$

where $h^S_{i,p}$ denotes the student prompt-boundary representation and $t^S_{i,p}$ is the reconstructed target from the token-coordinate transfer stage.The cosine objective aligns representation directions while remaining invariant to their magnitudes. The coefficient λ controls the representation-alignment weight.

4 Experiments

4.1 Experimental Setup

We evaluate SafetyTransfer across three teacher–student model pairs: Qwen3.5-9B → Qwen3.5-0.8B (Yang et al., 2025a), Qwen3-8B → Qwen3-0.6B (Yang et al., 2025a), and Llama-3.1-8B-Instruct → Llama-3.2-1B-Instruct (Grattafiori et al., 2024). The evaluation therefore covers two Qwen generations and the Llama3 family, with substantial differences in teacher and student model scale. We use four safety datasets: PAN (Pan et al., 2025), Safety-Tuned LLaMAs (STL) (Bianchi et al., 2024), WildGuardMix (WGM) (Han et al., 2024), and WildJailbreak (WJB) (Jiang et al., 2024).

We compare against the unadapted student (NoSFT), response-level supervised fine-tuning (SFT), and vanilla knowledge distillation (KD) (Hinton et al., 2015). Evaluation uses harmful refusal rate (HR), harmless over-refusal rate (OR), Gap ($\mathrm{HR} - \mathrm{OR}$), and SR-H, the harmonic mean of HR and harmless acceptance ($100 - \mathrm{OR}$). HR and OR separately measure refusal on the two input classes, while Gap and SR-H summarize whether high harmful refusal is achieved without excessive refusal of harmless requests. All reported methods follow the same evaluation and decoding protocol. Training, optimization, and implementation settings are provided in Appendix A.1.

Table 1: Selective-refusal performance of NoSFT, SFT, vanilla KD, and SafetyTransfer across three model pairs and four datasets. Avg. reports the mean across the four datasets. Bold and underline denote the best and second-best Gap and SR-H, respectively.

Data	Method	Qwen3.5				Qwen3				Llama3			
		HR↑	OR↓	Gap↑	SR-H↑	HR↑	OR↓	Gap↑	SR-H↑	HR↑	OR↓	Gap↑	SR-H↑
PAN	NoSFT	41.89	10.42	31.47	57.09	13.65	0.42	13.23	24.01	70.42	9.58	60.84	79.18
	SFT	97.92	32.08	65.83	80.20	94.79	18.13	<u>76.66</u>	<u>87.86</u>	97.71	15.62	82.09	90.56
	Vanilla KD	97.29	22.29	<u>75.00</u>	<u>86.40</u>	64.56	7.93	56.63	75.90	95.83	11.46	<u>84.37</u>	<u>92.04</u>
	Ours	97.71	8.13	**89.58**	**94.70**	97.71	10.83	**86.88**	**93.24**	97.29	7.08	**90.21**	**95.05**
STL	NoSFT	50.81	1.61	49.20	67.01	31.85	0.00	31.85	48.31	51.61	1.21	50.40	67.80
	SFT	87.10	4.84	82.26	<u>90.95</u>	81.45	1.61	<u>79.84</u>	<u>89.12</u>	80.24	1.21	<u>79.03</u>	<u>88.55</u>
	Vanilla KD	84.68	2.02	<u>82.66</u>	90.85	68.55	2.02	66.53	80.66	71.37	2.02	69.35	82.58
	Ours	89.92	3.63	**86.29**	**93.03**	91.53	3.63	**87.90**	**93.89**	89.11	2.02	**87.10**	**93.34**
WGM	NoSFT	55.97	18.62	37.35	<u>66.32</u>	23.61	4.13	19.48	37.89	80.37	21.06	59.31	79.65
	SFT	95.49	69.21	26.28	46.57	95.09	55.98	39.11	60.18	94.16	39.26	54.90	73.84
	Vanilla KD	93.90	53.76	<u>40.14</u>	61.97	72.94	16.72	<u>56.22</u>	<u>77.77</u>	85.68	16.63	<u>69.05</u>	<u>84.51</u>
	Ours	84.35	19.58	**64.77**	**82.34**	89.92	24.44	**65.48**	**82.11**	87.00	14.71	**72.29**	**86.14**
WJB	NoSFT	55.97	18.62	<u>37.35</u>	<u>66.32</u>	1.10	0.95	0.15	2.18	75.45	30.95	<u>44.50</u>	**72.11**
	SFT	94.55	61.43	33.12	54.79	95.45	86.67	8.78	23.39	96.20	57.14	39.06	59.30
	Vanilla KD	97.95	84.29	13.66	27.08	62.25	40.48	<u>21.77</u>	**60.85**	92.60	48.10	<u>44.50</u>	66.52
	Ours	86.65	36.19	**50.46**	**73.50**	81.40	54.76	**26.64**	<u>58.16</u>	87.10	40.95	**46.15**	<u>70.38</u>
Avg.	NoSFT	51.16	12.32	38.84	64.19	17.55	1.38	16.18	28.10	69.46	15.70	53.76	74.68
	SFT	93.76	41.89	51.87	<u>68.13</u>	91.70	40.60	<u>51.10</u>	65.14	92.08	28.31	63.77	78.06
	Vanilla KD	93.46	40.59	<u>52.87</u>	66.57	67.08	16.79	50.29	<u>73.80</u>	86.37	19.55	<u>66.82</u>	<u>81.41</u>
	Ours	89.66	16.88	**72.78**	**85.89**	90.14	23.42	**66.72**	**81.85**	90.13	16.19	**73.94**	**86.23**

4.2 Main Results

Table 1 compares SafetyTransfer with the unadapted student, SFT, and vanilla KD. SafetyTransfer achieves the highest Gap in all 12 model–dataset settings and the highest SR-H in 10. Averaged across datasets, its Gap exceeds the strongest comparison by 19.91, 15.62, and 7.12 points for Qwen3.5, Qwen3, and Llama3, respectively. The improvement therefore persists across all three teacher–student pairs rather than being concentrated in one model family. The aggregate results reveal a consistent change in the refusal operating point. Relative to SFT, SafetyTransfer has slightly lower average HR for Qwen3.5, Qwen3, and Llama3, but reduces average OR from 41.89 to 16.88, from 40.60 to 23.42, and from 28.31 to 16.19, respectively. The resulting Gap improvements over SFT are 20.91, 15.62, and 10.17 points. SafetyTransfer therefore improves selective refusal primarily by reducing responses that reject harmless inputs, rather than by uniformly increasing refusal.

This pattern is particularly clear on WGM. For Qwen3.5, SFT reaches 95.49 HR but also 69.21 OR, whereas SafetyTransfer obtains 84.35 HR with 19.58 OR, increasing Gap from 26.28 to 64.77 and SR-H from 46.57 to 82.34. Vanilla KD also attains high harmful refusal in several settings but can retain high over-refusal, as seen for Qwen3.5 on WJB, where 97.95 HR is accompanied by 84.29 OR. These results illustrate why harmful refusal alone does not characterize selective safety behavior. On WJB, SafetyTransfer retains the highest Gap for all three model pairs despite two settings in which another method attains a higher SR-H.

The magnitude and source of the improvement vary across datasets. On WGM, SFT already attains high harmful refusal but also exhibits substantial over-refusal, so the gain from SafetyTransfer is dominated by lower OR. In contrast, STL has low OR across the adapted models, and the improvement is driven more strongly by higher HR. PAN shows a similar but less pronounced balance between the two effects. This pattern is consistent across model families: compared with SFT, SafetyTransfer reduces average OR by 25.01, 17.18, and 12.12 points for Qwen3.5, Qwen3, and Llama3, while average HR changes by only 4.10, 1.56, and 1.95 points, respectively. The result-

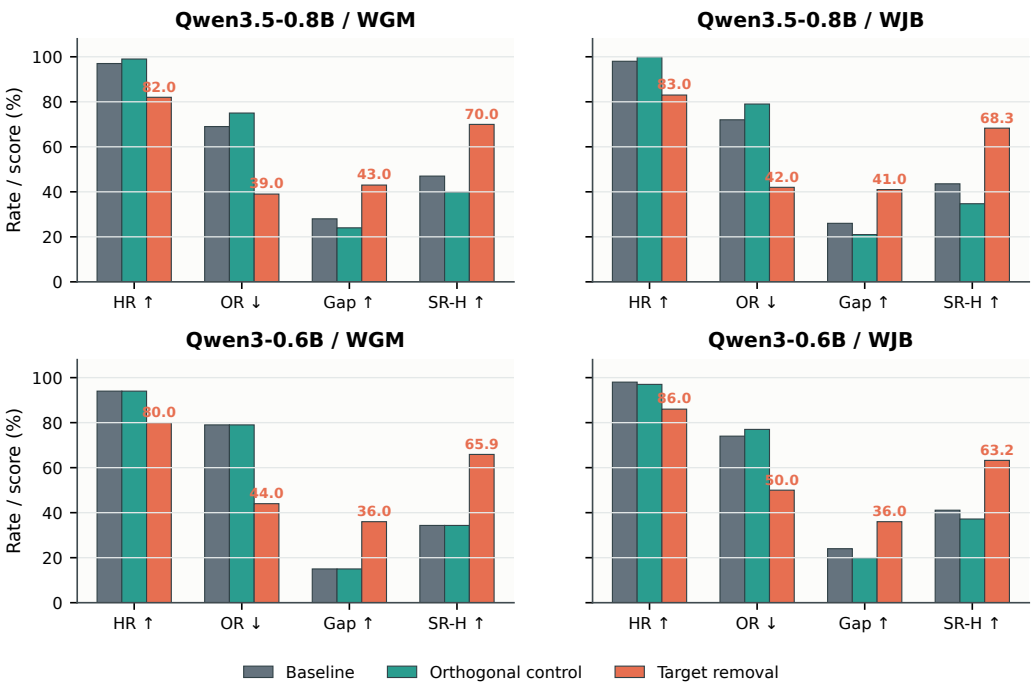

Figure 2: Effect of target-subspace intervention on selective refusal. Across two students and two datasets, 25% target-subspace removal produces a consistent behavioral shift, while a norm-matched orthogonal intervention remains close to standard inference.

ing gains therefore reflect a systematic change in the harmful-refusal–harmless-acceptance trade-off rather than a uniform increase in refusal.

4.3 Functional Role of the Transferred Target Subspace

We test the inference-time role of the transferred representations by intervening directly on the student target subspace. For Qwen3.5-0.8B and Qwen3-0.6B on WGM and WJB, we remove 25% of the last prompt-token projection onto the target subspace at each paired student layer and compare it with an equal-norm displacement in an orthogonal subspace. The intervention experiments use independently trained single-seed mechanism checkpoints and a fixed panel of 100 harmful and 100 harmless prompts per dataset, whereas Table 1 reports the main benchmark evaluation. Accordingly, the no-intervention points in Figures 2 and 3 and Table 4 serve as within-run references for the paired intervention conditions rather than repeated measurements of Table 1. All intervention contrasts are computed within the same checkpoint, prompt panel, and decoding configuration.

Targeted subspace removal. Figure 2 shows that orthogonal perturbations remain close to standard inference, whereas target-subspace removal consistently shifts refusal behavior. Relative to the orthogonal control, targeted removal produces a substantial change in both HR and OR. Importantly, the intervention does not affect harmful refusal in isolation: refusal decreases for both harmful and harmless inputs. The transferred directions therefore participate in the student's refusal operating point rather than affecting only one input class. Because the decrease in OR is larger than the decrease in HR in the evaluated settings, the intervention moves the model toward greater harmless acceptance while simultaneously reducing harmful refusal.

Method specificity. We next apply the same intervention to SafetyTransfer, SFT, and random-target checkpoints. Figure 3 shows the largest target-removal displacement for SafetyTransfer across all four student–dataset settings, while the corresponding orthogonal perturbations remain comparatively small. SFT and random-target checkpoints show smaller responses to the same targeted operation. This difference links the representation-alignment objective to increased sensitivity along the

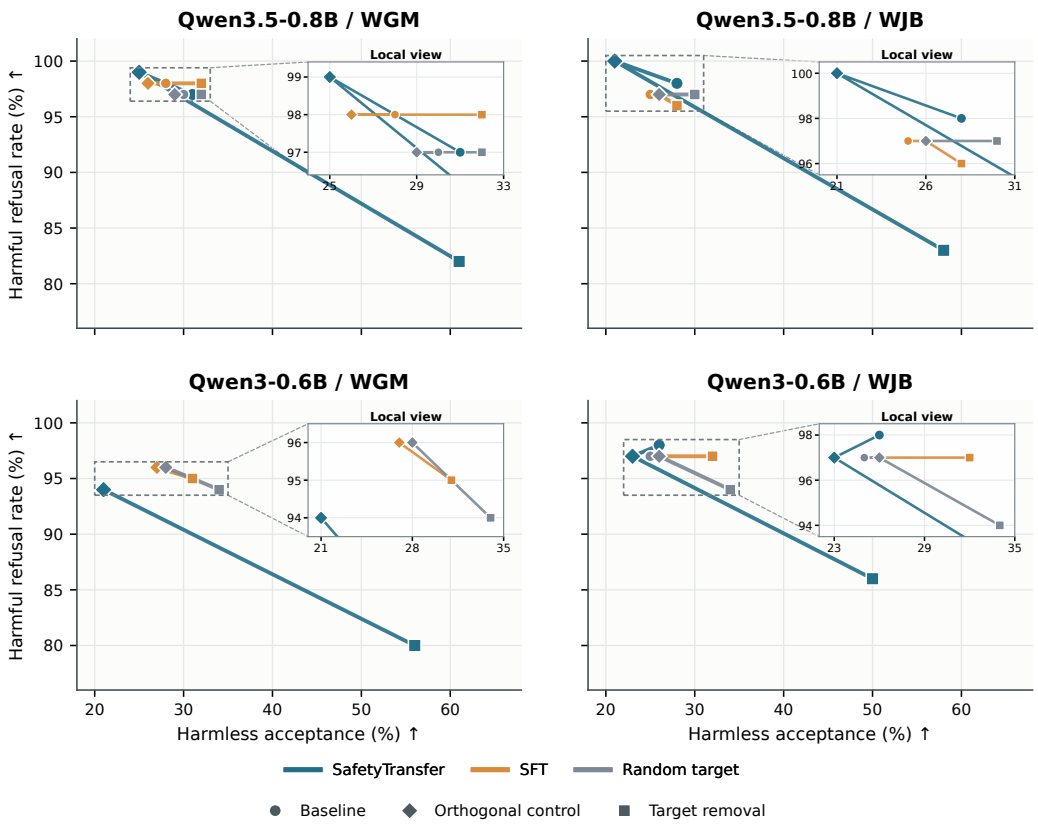

Figure 3: Method-specific effect of target-subspace removal. Across SafetyTransfer, SFT, and random-target training, targeted removal produces the largest shift in the harmful-refusal–harmless-acceptance operating point for SafetyTransfer, while norm-matched orthogonal perturbations remain close to standard inference.

transferred target directions. Together, the two intervention analyses establish a functional connection between the learned representation geometry and inference-time refusal behavior.

The direction of the intervention provides additional information about the transferred representation. Target removal shifts both harmful and harmless inputs toward lower refusal, rather than changing only harmful-request behavior. The transferred directions therefore participate in the broader refusal operating point, with an asymmetric effect across the two input classes. In the evaluated settings, OR changes more strongly than HR, which explains why removing the target projection can increase Gap even as harmful refusal decreases. Together with the small orthogonal-control shifts, this pattern shows that the observed effect depends on the geometry of the transferred directions rather than perturbation magnitude alone.

4.4 Ablation Studies

Multi-dimensional contrastive representation. We replace the proposed low-rank contrastive subspace with a single normalized harmful-minus-harmless mean direction while keeping the remaining transfer procedure fixed. Both variants are evaluated on PAN and WGM across all three teacher–student pairs. The contrastive subspace achieves higher Gap and SR-H in all six settings. On PAN, Gap improves by 3.12–3.75 points and SR-H by 1.58–2.12 points. The difference is larger on WGM, where the subspace reduces OR by 13.65–24.97 points and improves Gap by 10.99–20.06 points. Notably, the mean-direction variant often retains comparable or even higher HR on WGM, but this is accompanied by substantially higher OR. For Qwen3.5, for example, the mean direction raises HR from 84.35 to 89.26 while increasing OR from 19.58 to 44.55. The advantage of the

Table 2: Effect of contrastive dimensionality on selective refusal. SafetyTransfer uses the low-rank contrastive subspace, while Mean direction uses a single harmful-minus-harmless activation direction with all other transfer components fixed.

		ST (Subspace)				Mean direction			
Data	Model pair	HR↑	OR↓	Gap↑	SR-H↑	HR↑	OR↓	Gap↑	SR-H↑
PAN	Qwen3.5 9B→0.8B	97.71	8.13	**89.58**	**94.70**	96.46	10.00	86.46	93.12
PAN	Qwen3 8B→0.6B	97.71	10.83	**86.88**	**93.24**	97.29	13.75	83.54	91.44
PAN	Llama3 8B→1B	97.29	7.08	**90.21**	**95.05**	98.54	12.08	86.46	92.93
WGM	Qwen3.5 9B→0.8B	84.35	19.58	**64.77**	**82.34**	89.26	44.55	44.71	68.40
WGM	Qwen3 8B→0.6B	89.92	24.44	**65.48**	**82.11**	89.92	40.63	49.29	71.52
WGM	Llama3 8B→1B	87.00	14.71	**72.29**	**86.14**	89.66	28.36	61.30	79.64

low-rank representation therefore comes from a better separation of harmful refusal from harmless acceptance, rather than simply producing a stronger refusal signal.

Representation alignment under shared response supervision. We compare SafetyTransfer with GKD and DistiLLM-2 under the common objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + 0.25\mathcal{L}_{\mathrm{aux}},$$

where the methods differ in their auxiliary transfer objective. For Qwen3-8B→Qwen3-0.6B on PAN and STL, SafetyTransfer achieves the highest Gap and SR-H on both datasets. Averaged across them, Gap exceeds GKD by 5.00 points and DistiLLM-2 by 5.62 points, with corresponding SR-H gains of 2.75 and 3.11 points.

Table 3: SafetyTransfer, GKD, and DistiLLM-2 under shared response supervision. All methods use the same cross-entropy objective and differ in their auxiliary transfer signal.

Data	Method	HR↑	OR↓	Gap↑	SR-H↑
PAN	ST	97.71	10.83	**86.88**	**93.24**
	GKD	97.29	13.96	83.33	91.32
	DistiLLM-2	97.50	15.00	82.50	90.82
STL	ST	91.53	3.63	**87.90**	**93.89**
	GKD	84.68	3.23	81.45	90.32
	DistiLLM-2	84.27	3.23	81.05	90.09
Avg.	ST	94.62	7.23	**87.39**	**93.57**
	GKD	90.98	8.59	82.39	90.82
	DistiLLM-2	90.89	9.11	81.77	90.46

The two datasets show different sources of this improvement. On PAN, SafetyTransfer maintains similar HR while reducing OR relative to both distillation baselines. On STL, OR remains near 3% for all three methods, while SafetyTransfer raises HR from approximately 84% to 91.53%. Representation alignment therefore improves the selective-refusal balance through different sides of the operating point depending on the dataset, rather than acting as a uniform increase or decrease in refusal. Taken together, the two ablations isolate complementary parts of SafetyTransfer. The contrastive-subspace ablation concerns the representation extracted from the teacher, while the shared-CE comparison concerns how the transferred signal supervises the student. The first result shows that retaining multiple contrastive directions is most beneficial when a single mean direction produces high refusal on both harmful and harmless inputs. The second shows that representation alignment remains advantageous when response supervision is held fixed, with the source of the gain depending on the dataset: lower OR on PAN and higher HR on STL.

5 Conclusion

We introduce **SafetyTransfer**, a representation-level framework for transferring harmful–harmless discrimination from aligned teachers to compact students. SafetyTransfer extracts low-rank contrastive components from informative teacher representations, reconstructs them across hidden dimensions through a token-semantic interface, and combines representation alignment with response supervision. Across three teacher–student pairs and four safety datasets, SafetyTransfer achieves the highest Gap among the main baselines and the highest SR-H in 10 settings. The results consistently show that its gains arise from a better balance between harmful refusal and harmless acceptance rather than uniformly stronger refusal. Targeted interventions connect the transferred directions to inference-time refusal behavior, while the ablations separately support multi-dimensional contrastive extraction and representation-level supervision.

AI USE STATEMENT

Generative AI tools were used to assist manuscript drafting and language polishing, and to support the discovery and retrieval of related literature. All AI-assisted content was reviewed and revised by the authors. The authors take responsibility for the final text, analyses, and claims in this work.

ETHICS STATEMENT

This work uses public language-model safety benchmarks and does not involve human subjects or private data collection. All experiments are conducted on language models under controlled benchmark settings. The released materials are intended to support reproducible research and further understanding of model safety and behavior.

REPRODUCIBILITY STATEMENT

We provide code and implementation details to facilitate reproduction of our experiments. The main paper specifies the representation extraction, contrastive component construction, token-semantic transfer, training objective, evaluation metrics, and intervention protocol. Appendix A.1 provides implementation details for layer pairing, component construction, token-semantic reconstruction, optimization, and evaluation. Exact values for the target-subspace intervention are also reported in the appendix. The complete source code is provided in the supplementary materials.

REFERENCES

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, 2024. URL https://arxiv.org/abs/2406.11717.

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.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AAXMcAyNF6.

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Tung Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep, and Trung Le. Sra: Span representation alignment for large language model distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32961–32975, 2026.

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.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. Vocabulary-defined semantics: Latent space clustering for improving in-context learning. *arXiv preprint arXiv:2401.16184*, 2024a.

Jian Gu, Aldeida Aleti, Chunyang Chen, and Hongyu Zhang. Beyond neural incompatibility: Cross-scale knowledge transfer in language models through latent semantic alignment, 2026. URL https://arxiv.org/abs/2510.24208.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024b.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.185. URL https://aclanthology.org/2025.acl-long.185/.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Thinh On, Hengzhi Pei, Leonard Lausen, and George Karypis. Knowledge distillation for large language models through residual learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

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.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47697–47716. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pan25f.html.

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. Safety subspaces are not linearly distinct: A fine-tuning case study. In *International Conference on Learning Representations*, volume 2026, pp. 95677–95710, 2026.

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.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL https://arxiv.org/abs/1412.6550.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, and Jie Tang. GKD: A general knowledge distillation framework for large-scale pre-trained language model. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 134–148, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.15. URL https://aclanthology.org/2023.acl-industry.15/.

Mykola Vysotskyi, Zahar Kohut, Anna-Alina Bondarets, Taras Rumezhak, and Volodymyr Karpiv. Cross-modal representational alignment with LLM priors for image generation. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025. URL https://openreview.net/forum?id=SQQWxwlXWF.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety mechanism transfer in large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 58315–58336, 2025.

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

Junjie Yang, Junhao Song, Xudong Han, Ziqian Bi, Tianyang Wang, Chia Xin Liang, Xinyuan Song, Yichao Zhang, Qian Niu, Benji Peng, et al. Feature alignment and representation transfer in knowledge distillation for large language models. *arXiv preprint arXiv:2504.13825*, 2025b.

Jingyu Zhang, Ahmed Elgohary Ghoneim, Ahmed Magooda, Daniel Khashabi, and Ben Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. In *International Conference on Learning Representations*, volume 2025, pp. 68007–68038, 2025.

A Appendix

A.1 Implementation Details

Representation extraction and layer pairing. Teacher and student representations are extracted using each model's native chat template with the assistant generation prompt. We perform a forward pass without generation and use the hidden state at the last non-padding prompt position, corresponding to the state that predicts the first assistant token. The embedding output is excluded. Selected teacher layers are paired with student layers by relative depth:

$$\ell_S = \mathrm{clamp}\left(\left\lfloor \frac{\ell_T + 1}{L_T} L_S \right\rfloor, 1, L_S\right) - 1. \tag{6}$$

When multiple teacher layers map to the same student layer, their alignment terms are retained independently.

Teacher component construction. Layer selection combines geometric harmful–harmless separation with the validation balanced accuracy of a linear probe trained on standardized representations using LBFGS. Before contrastive subspace construction, hidden states are L2-normalized and harmful and harmless samples are balanced by deterministic sampling. The retained rank is the smallest rank satisfying the specified singular-value energy threshold subject to an upper-rank constraint. Contrastive directions are estimated from normalized states, while the original hidden states are used for projection.

Token-semantic reconstruction. The token-semantic interface assumes compatible token identities between teacher and student models. The Gram matrix $W^\top W$ and its pseudoinverse are computed in FP32, with large vocabulary matrices processed in chunks. Teacher components are normalized before semantic decomposition. We retain the top-K coefficients by absolute magnitude while preserving their signs, after filtering special, empty, and formatting-only tokens.

Optimization and evaluation. The teacher remains frozen and students are optimized with LoRA adapters of rank 16, alpha 32, and dropout 0.05, applied to attention and MLP projection modules. We use AdamW with coefficients $(0.9, 0.95)$, cosine learning-rate decay, 3% warmup, zero weight decay, and gradient clipping at 1.0. The maximum training sequence length is $2{,}048$ tokens. Evaluation uses deterministic decoding with maximum prompt and generation lengths of $4{,}096$ and $1{,}024$ tokens, respectively. At inference, only the trained student model is required.

A.2 Exact Values for the Target-Subspace Intervention

Table 4 reports the exact paired values underlying Figure 2. These values are obtained from the single-seed mechanism checkpoints and fixed intervention panels described in Section 4.3.

A.3 Robustness to Component and Target Variants

We further evaluate the sensitivity of SafetyTransfer to three component-level changes on Llama-3.1-8B-Instruct→Llama-3.2-1B-Instruct across PAN, STL, WGM, and WJB. *NoProj* bypasses the contrastive-subspace projection before token-semantic reconstruction. *WithinLabel* constructs the transfer target within the same safety label, while *CrossLabel* exchanges the label-conditioned target assignment. All variants use the same optimization and evaluation protocol as the complete method. The SafetyTransfer results are reproduced from the main comparison for reference.

Across the four datasets, the complete SafetyTransfer configuration achieves the highest average Gap of 73.94 and SR-H of 86.23. Relative to NoProj, WithinLabel, and CrossLabel, its average Gap is higher by 0.82, 0.84, and 1.65 points, respectively, with corresponding SR-H gains of 0.61, 0.59, and 0.95 points. The separation is most pronounced on WGM, where SafetyTransfer reduces OR to 14.71 and improves Gap by 2.11–3.99 points and SR-H by 1.15–2.24 points over the three variants. Differences are smaller on the remaining datasets, where individual variants can closely match the complete configuration. Overall, these results show that the complete design provides the strongest aggregate selective-refusal performance, with its clearest advantage emerging on the more challenging WGM distribution.

Table 4: Exact selective-refusal metrics for the intervention in Figure 2. All values are percentages. Gap is defined as $\mathrm{HR} - \mathrm{OR}$, and SR-H is the harmonic mean of HR and harmless acceptance $(100 - \mathrm{OR})$.

Student	Dataset	Condition	HR↑	OR↓	Gap↑	SR-H↑
Qwen3.5-0.8B	WGM	Standard inference	97.0	69.0	28.0	47.0
		Orthogonal control	99.0	75.0	24.0	39.9
		25% target removal	82.0	39.0	43.0	70.0
	WJB	Standard inference	98.0	72.0	26.0	43.6
		Orthogonal control	100.0	79.0	21.0	34.7
		25% target removal	83.0	42.0	41.0	68.3
Qwen3-0.6B	WGM	Standard inference	94.0	79.0	15.0	34.3
		Orthogonal control	94.0	79.0	15.0	34.3
		25% target removal	80.0	44.0	36.0	65.9
	WJB	Standard inference	98.0	74.0	24.0	41.1
		Orthogonal control	97.0	77.0	20.0	37.2
		25% target removal	86.0	50.0	36.0	63.2

Table 5: Component-level variants of SafetyTransfer on Llama-3.1-8B-Instruct→Llama-3.2-1B-Instruct. Avg. denotes the arithmetic mean across the four datasets. All values are percentages.

Data	Variant	HR↑	OR↓	Gap↑	SR-H↑
PAN	NoProj	98.12	7.29	90.83	95.34
	WithinLabel	97.92	6.88	91.04	95.46
	CrossLabel	97.92	9.58	88.33	94.02
	SafetyTransfer	97.29	7.08	90.21	95.05
STL	NoProj	87.90	1.61	86.29	92.85
	WithinLabel	87.90	2.02	85.89	92.67
	CrossLabel	89.11	2.82	86.29	92.99
	SafetyTransfer	89.11	2.02	87.10	93.34
WGM	NoProj	88.06	17.88	70.18	84.99
	WithinLabel	88.59	19.58	69.02	84.31
	CrossLabel	88.73	20.42	68.30	83.90
	SafetyTransfer	87.00	14.71	72.29	86.14
WJB	NoProj	88.05	42.86	45.19	69.31
	WithinLabel	88.35	41.90	46.45	70.10
	CrossLabel	87.65	41.43	46.22	70.22
	SafetyTransfer	87.10	40.95	46.15	70.38
Avg.	NoProj	90.53	17.41	73.12	85.62
	WithinLabel	90.69	17.60	73.10	85.64
	CrossLabel	90.85	18.56	72.29	85.28
	SafetyTransfer	90.13	**16.19**	**73.94**	**86.23**