Understanding the Role of Prompt Template in Knowledge Distillation for Safety Alignment

Anjila Budathoki[1], Manish Dhakal[1], Benjamin M. Ampel[2], Yi Ding[1]
[1]University of Tennessee, Knoxville
[2]Georgia State University
{abudatho, mdhakal1}@vols.utk.edu, bampel@gsu.edu, yding@utk.edu

Abstract

Prior research has demonstrated that the choice of prompt template during Supervised Fine-Tuning (SFT) significantly impacts the robustness of safety alignment afterwards. However, the influence of template selection during Knowledge Distillation (KD) from teacher to student remains largely unexplored. Thus, we fill this gap by analyzing how different template configurations influence the pre-existing safety alignment of the student. We observe a significant degradation of safety alignment present in the aligned base instruct-tuned model. Specifically, we find that utilizing chat templates renders the model more compliant with harmful queries compared to a non-chat template. These findings are consistent across three models: LLaMA, Gemma and Qwen model families and are evaluated across multiple safety benchmarks. We further show that using a non-chat template during distillation better preserves the base student's internal representations, while chat template distillation induces a larger representational shift.[1]

Warning: this paper includes examples that may be offensive or harmful.

1 Introduction

Knowledge Distillation (KD) (Hinton et al., 2015) is an effective technique for compressing large models into smaller, efficient student models while retaining performance. In the context of Large Language Models (LLMs), KD is commonly used to obtain compact models that are easier to deploy under computational constraints. As these models are increasingly integrated into practical applications such as healthcare and autonomous systems (MohiEldeen Alabbasy et al., 2023; Agand, 2024), considerations of robustness and safety become increasingly important. In particular, deployed distilled models are expected to exhibit safe and harmless behavior, including refusing harmful, malicious, or policy-violating requests. To encourage such behavior, modern instruction-tuned models typically undergo alignment processes that embed refusal capabilities before any downstream adaptation (Askell et al., 2021; Ouyang et al., 2022). A key open question is how the safety alignment of student models changes during further training via KD, and which training-time factors govern its preservation or degradation.

In particular, the choice of prompt template (Lyu et al., 2024) during training is one such factor. Prompt templates specify how inputs are formatted for the model, often as strings with placeholders filled by user queries, instructions, and assistant responses. Modern LLMs (Touvron et al., 2023; Jiang et al., 2023) primarily utilize *chat templates*, which standardize user interactions via specific control tokens (e.g., `<|user|>`, `[INST]`); we contrast these with *non-chat templates*, which present the same content without these conversational control tokens. Hereafter, we use *chat templates* to refer to templates that wrap inputs with special conversational control tokens, and *non-chat templates* to refer to task-style formats that omit these tokens. Prior work shows that such formatting choices can substantially affect safety behavior during SFT (Lyu et al., 2024; Jiang et al., 2025; Wang et al., 2025b); however, their role in KD remains empirically underexplored.

Unlike SFT, KD exposes the student to teacher-generated soft targets, an additional signal that may interact with template formatting in ways that distinctly affect the student's safety representations. This raises an important question about the interplay between distillation, template formatting, and safety retention:

How do chat and non-chat prompt templates affect the preservation of safety-aligned refusal behavior when student models are distilled on benign downstream tasks?

[1]Code:https://github.com/anjilab/role-of-prompt-template-in-kd

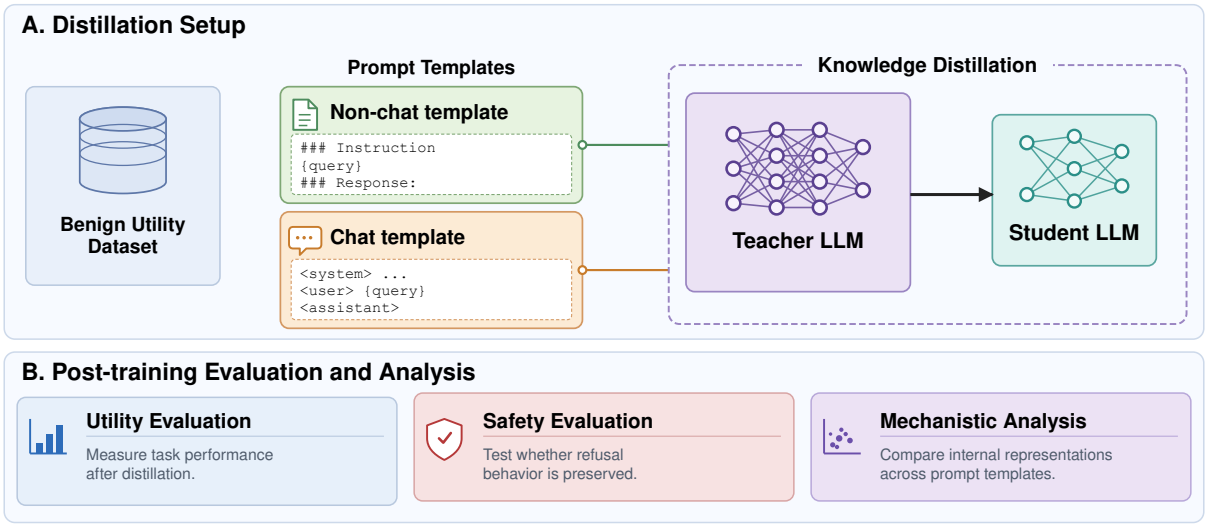

Figure 1: Overview of the experimental pipeline. We distill student models on benign instruction-following data under two formatting conditions: non-chat and chat. The pipeline compares how these training-time prompt templates affect the resulting student model during downstream safety and utility evaluation.

In this work, we answer this question by systematically investigating how the format of prompt templates during training shapes the distilled student model's susceptibility to harmful queries (Qi et al., 2025; Arditi et al., 2024). We distill student models on benign instruction-following data under two controlled template conditions: chat and non-chat, and evaluate all models using the standard chat template at inference, ensuring that any observed safety differences reflect training-time choices alone. Beyond output-level evaluation, we conduct a mechanistic analysis to examine whether chat-template KD shifts the student's internal refusal representations or whether the behavioral gap is a surface-level artifact. We further study *prompt template mixing* by varying the proportion of chat versus non-chat samples during training to characterize how safety and utility scale with chat template exposure. The overall pipeline is shown in Figure 1.

Our key findings are summarized as follows:

- Knowledge distillation on benign dataset can erode the pre-existing safety alignment of aligned student models across model families.
- Prompt formatting modulates the severity of this degradation: using a chat template during KD consistently leads to a higher Attack Success Rate (ASR) than non-chat KD under identical training data.
- The student-side training template, not the teacher's output distribution, is the primary driver: using the chat template only for the teacher causes minimal safety loss, whereas using it for the student produces consistent regression.
- The degradation is not limited to output behavior: chat-template KD shifts the student's internal refusal direction away from the base model and reduces its ability to separate harmful from harmless prompts, whereas non-chat KD largely preserves both.
- Safety is more sensitive to chat template exposure than utility. As the proportion of chat template samples increases, ASR generally rises, indicating amplified safety degradation, whereas utility improves only modestly.

2 Related work

Knowledge Distillation. Hinton et al. (2015) originally demonstrated that soft targets (i.e., a full probability distribution) encode richer information than hard labels (i.e., a discrete label). This enabled the efficient transfer of knowledge from a large deep neural network to smaller ones. Recently, research has shifted towards distilling the knowledge of LLMs into more compact student models (Gu et al., 2024; Ko et al., 2024; Wang et al., 2025a). Notably, Gu et al. (2024) introduced a distillation objective for generative models (adopted by many subsequent methods) which introduced reverse KL divergencefor improved KD in generative LLMs. However, despite the advancement in task-centric

KD that prioritizes task utility (e.g., preserving accuracy or fluency), how these processes impact safety alignment is currently under-studied.

Safety Alignment in LLMs. To align LLMs with human values, models often adopt a two-stage post-training pipeline of SFT and alignment-tuning (e.g., Reinforcement Learning from Human Feedback, Direct Preference Optimization) (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). These methods have demonstrated that careful alignment can embed the ability to reject harmful instructions (known as safety guardrails) within an LLM. Prior work has also employed KD as a mechanism to enforce safety alignment in student models Yang et al. (2024). However, literature has found that several vulnerabilities exist in the safety guardrails of open-access LLMs (Yi et al., 2024). These vulnerabilities lead to jailbreak attacks, which allow malicious users to use the LLM in unintended ways (Yi et al., 2024). It is unclear if distilling models on benign downstream tasks can further degrade these safety guardrails in previously aligned models.

Prompt Templates. Prompt templates, which determine how instructions, user inputs, and assistant responses are serialized before being passed to a model, are often treated as implementation details, but these formatting choices meaningfully shape model behavior and safety. During instruction fine-tuning, chat templates can reduce context awareness, influencing how models attend to input (Wang et al., 2025b). They also affect safety: task-style fine-tuning and safety-oriented testing better preserve safe behavior (Lyu et al., 2024), while certain chat template designs induce unexpected behavioral failures (Jiang et al., 2025). Similar vulnerabilities also appear in VLMs, where Role-Modality Attacks exploit dialogue roles and modality placement (Shayegani et al., 2026). Despite these findings, the role of prompt templates in knowledge distillation remains underexplored; we address this gap by systematically comparing chat and non-chat templates during benign KD and measuring their impact on both utility and safety.

3 Methodology

In this section, we present our approach for analyzing the impact of prompt templates on safety alignment during Knowledge Distillation (KD). Our pipeline, illustrated in Figure 1, consists of two stages: (1) constructing instruction datasets using distinct prompt formatting strategies (chat vs. non-chat), and (2) fine-tuning a student model using a balanced KD objective.

3.1 Prompt Formatting

To examine the role of structural cues during distillation, we design two controlled formatting conditions for the instruction dataset $\mathcal{D}$. While modern instruction-tuned models typically require specific chat templates (e.g., <|begin_of_text|>, <|start_header_id|>) to maintain state and role (Hugging Face Team, 2025), it is unclear if these tokens aid or hinder the transfer of safety representations during KD.

We use standard definition of two template configurations, as described below (see e.g. in Table 9):

- **Chat Template:** We use each model family's native chat template, including all special control tokens, as the chat-format baseline for instruction following.
- **Non-chat Template:** We strip all model-specific control tokens, presenting the input as raw text. This isolates the semantic content of the instruction from the structural priors enforced by the template.

Importantly, regardless of the training template, all models are evaluated using the standard chat template at inference to reflect real-world deployment conditions (details in §3.3.3)

3.2 Distillation Objective

Given the formatted dataset $\mathcal{D} = \{(x, y)\}$, we fine-tune the student model q_θ by combining a standard supervised learning objective with Knowledge Distillation from the teacher p. The total loss function balances the ground-truth alignment with the transfer of the teacher's probability distribution:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{KD}$$

The supervised component, $\mathcal{L}_{CE}$, enforces alignment with the gold reference tokens, while $\mathcal{L}_{KD}$ minimizes the forward KL divergence between the teacher and student logits:

$$\mathcal{L}_{CE} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\log q_\theta(y|x)\right]$$
$$\mathcal{L}_{KD} = \mathbb{E}_{x\sim\mathcal{D},y\sim p(\cdot|x)}\left[\log p(y|x) - \log q_\theta(y|x)\right]$$

We set $\alpha = 0.5$ to assign equal weight to task performance and knowledge transfer. Although our

primary experiments use forward KL as the standard distillation objective, we additionally evaluate reverse KL and combined FKL+RKL objectives to assess whether the observed template effects generalize across distillation objectives (Appendix C).

3.3 Experimental Setup

In this section, we detail our experimental setup for investigating how prompt template choice during distillation influences the preservation of safety alignment of the student models.

3.3.1 Models.

We employ three widely used open-weight model families: LLaMA-3 (Grattafiori et al., 2024), Gemma-2 (Team et al., 2024), and Qwen-2.5 (Qwen et al., 2024). Following standard KD protocols (Gu et al., 2024; Ko et al., 2024), we use larger Instruct variants as **teachers** ($\mathcal{M}_T$) and smaller Instruct variants as *students* ($\mathcal{M}_S$): **LLaMA-3.1-8B-Instruct** $\rightarrow$ *LLaMA-3.2-3B-Instruct*, **Gemma-2-9B-IT** $\rightarrow$ *Gemma-2-2B-IT*, and **Qwen2.5-7B-Instruct** $\rightarrow$ *Qwen2.5-3B-Instruct*. These model pairs were selected due to their rigorous pretraining filtration and post-training safety alignment, providing a strong baseline for measuring safety degradation.

3.3.2 Datasets.

Instruction Following. We perform distillation with the `databricks-dolly-15k` dataset (Gu et al., 2024), a standard benchmark for instruction following.

Safety Evaluation. To measure safety alignment, we evaluate on four adversarial benchmarks: AdvBench (Zou et al., 2023) and JailbreakBench (Chao et al., 2024) for harmful-instruction attacks, and HarmBench (Mazeika et al., 2024) and SORRY-Bench (Xie et al., 2025) for broader harmful-behavior coverage. These datasets are used for evaluation only; please refer to Appendix A.4 for a detailed description of each.

3.3.3 Training and Evaluation

Training We fine-tune all student models using LoRA (Hu et al., 2022) rank $r = 8$. The training is conducted on a single node equipped with 4 $\times$ NVIDIA RTX 4090 GPUs. For a comprehensive overview of hyperparameters, including learning rates and batch sizes, please refer to the Appendix A.1.

Evaluation Setting. We assess model performance along two dimensions: task utility and safety alignment. For task utility, we report the average ROUGE-L score across the four general-domain instruction-following sets. For safety alignment, we compute Attack Success Rate (ASR) on AdvBench and JailbreakBench using the WalledEval framework (Gupta et al., 2024) with `LLaMA-3-Guard-8B` as the judge, and evaluate HarmBench and SORRY-Bench using their respective provided judges.

Inference Setting. Crucially, during evaluation, we strictly adhere to the standard chat template for all models (including those trained with non-chat templates). This follows the recommended inference-time usage of modern instruction-tuned LLMs (Grattafiori et al., 2024; Team et al., 2024) and ensures that our evaluation reflects the practical safety of deployed models. Therefore, any observed safety differences across conditions can be attributed to the training-time template choice, rather than to differences in the evaluation format.

4 Results

We analyze the safety behavior of student models after knowledge distillation (KD) under two training-template conditions: (1) chat template and (2) non-chat template. During evaluation, all models are tested using the standard chat template, reflecting the expected deployment setting.

4.1 Distillation generally erodes safety.

We observe that standard KD on benign instruction-following data degrades the safety alignment in the student models. As shown in Table 1, all three distilled student models show increased ASR on most safety benchmarks relative to their corresponding base student models. In some cases, the vulnerability of the student models has almost tripled with maximal degradation (HarmBench for LLaMA and SORRY-Bench for Gemma). We further show in Appendix C that the same template-driven safety degradation persists under alternative distillation objectives.

4.2 Template Choice Shapes Safety Degradation

Our main finding is that the prompt template used during KD strongly affects safety preservation. When the student is distilled using the standard

Model	Method	Training Template	Safety (ASR% ↓)				Utility (RL↑)
			AdvBench	JBB	SORRY	HarmBench	AVG
LLaMA-3.2-3B-Instruct	$\mathcal{M}_S$	-	2.98	7.25	26.72	12.18	19.84
	SFT	Non-chat	$0.0096_{-2.97}$	$0.04_{-7.21}$	$27.12_{+0.40}$	$12.19_{+0.01}$	24.27
	SFT	Chat	$0.04_{-2.94}$	$0.07_{-7.18}$	$28.71_{+1.99}$	$22.50_{+10.32}$	29.10
	KD	Non-chat	$3.07_{+0.09}$	$6.00_{-1.25}$	$27.36_{+0.64}$	$29.81_{+17.63}$	24.75
	KD	Chat	$\mathbf{4.99}_{+2.01}$	$\mathbf{8.75}_{+1.50}$	$\mathbf{30.86}_{+4.14}$	$\mathbf{39.75}_{+27.57}$	29.03
Gemma-2-2B-IT	$\mathcal{M}_S$	-	0.00	1.50	16.04	11.56	22.35
	SFT	Non-chat	$0.00_{+0.00}$	$0.015_{-1.485}$	$28.71_{+12.67}$	$20.37_{+8.81}$	23.66
	SFT	Chat	$0.028_{+0.028}$	$0.055_{-1.445}$	$53.18_{+37.14}$	$13.25_{+1.69}$	30.05
	KD	Non-chat	$0.00_{+0.00}$	$1.50_{+0.00}$	$19.18_{+3.14}$	$14.25_{+2.69}$	23.63
	KD	Chat	$\mathbf{4.00}_{+4.00}$	$\mathbf{6.50}_{+5.00}$	$\mathbf{51.36}_{+35.32}$	$\mathbf{19.56}_{+8.00}$	30.18
Qwen-2.5-3B-Instruct	$\mathcal{M}_S$	-	0	2.75	36.29	17.69	22.46
	SFT	Non-chat	$0.67_{+0.67}$	$3.25_{+0.50}$	$35.15_{-1.14}$	$17.75_{+0.06}$	23.93
	SFT	Chat	$4.90_{+4.90}$	$12.75_{+10.00}$	$41.67_{+5.38}$	$31.94_{+14.25}$	29.15
	KD	Non-chat	$1.06_{+1.06}$	$3.25_{+0.50}$	$34.93_{-1.36}$	$17.81_{+0.12}$	23.96
	KD	Chat	$\mathbf{5.87}_{+5.87}$	$\mathbf{11}_{+8.25}$	$\mathbf{40.99}_{+4.70}$	$\mathbf{30.81}_{+13.12}$	28.95

Table 1: Student-model safety and utility under different prompt templates. We evaluate how prompt templates used during knowledge distillation affect the resulting student model's safety and utility. Safety is measured using Attack Success Rate (ASR%), where lower values indicate safer behavior. Utility reports the average ROUGE-L score across four instruction-following benchmarks: Dolly, SelfInst, S-NI, and Vicuna. The subscripted deltas show the change relative to the corresponding student baseline ($\mathcal{M}_S$) within each model family. Red deltas indicate increased ASR, corresponding to worse safety, while green deltas indicate reduced ASR, corresponding to improved safety. To emphasize the KD template effect, bold values are used only in the KD safety columns and denote the higher (less safe) ASR between chat and non-chat KD.

chat template, the resulting model shows larger increases in ASR than with the non-chat template. For example, on SORRY-Bench, the Gemma student shows a **+35.32%** ASR increase relative to its base student model. Similarly, on HarmBench, we observe a **+27.57%** increase for LLaMA and a **+13.12%** increase for Qwen. These results suggest that chat-template KD is associated with greater erosion of pre-existing refusal behavior.

While prior work has shown that SFT can degrade safety alignment (Lyu et al., 2024; Jiang et al., 2025), we show that similar template sensitivity also emerges in KD, even when the distillation data are benign. Importantly, the effect is substantially weaker with non-chat KD. On Gemma, SORRY-Bench ASR increases by only **+3.14%**, compared with **+35.32%** under chat-template KD. The same pattern holds across model families: LLaMA shows a smaller HarmBench increase, while Qwen remains near baseline on HarmBench (**+0.12%**) but degrades substantially under chat-template KD (**+13.12%**).

These results suggest that using a non-chat template during KD reduces safety degradation under standard chat-template evaluation. However, non-chat KD does not fully preserve the original alignment. For example, LLaMA still shows a **+17.63%** HarmBench ASR increase, indicating that distillation itself weakens refusal behavior. Although non-chat KD still improves utility over the base student model, its gains are smaller than those achieved by chat template KD.

4.3 The safety-utility trade-off.

Given that KD is primarily used to improve task utility, it is important to examine whether template choice also affects this dimension. Our results reveal a clear trade-off: KD improves instruction-following under both template conditions, but larger utility gains coincide with larger ASR increases. As shown in Table 1, chat-template KD achieves higher utility but also incurs the largest safety costs, whereas non-chat KD yields smaller utility gains while better preserving safety. Appendix E further shows that chat-template KD accelerates this vulnerability earlier in training.

One possible explanation is that the lower ASR of non-chat KD simply reflects weaker instruction-following rather than better safety preservation. However, this is inconsistent with the data: for LLaMA, non-chat KD achieves lower utility than SFT Chat (24.75 vs. 29.10) yet shows higher Harm-

Bench ASR (29.81% vs. 22.50%). If reduced compliance explained the lower ASR, lower utility should correspond to lower ASR, which is not observed. Section 5.2 further shows that non-chat KD better preserves refusal-related representations, suggesting deeper mechanistic differences between the two template conditions.

5 Analysis

The observed safety degradation under chat-template KD Prompts us to investigate three follow-up questions: first, whether this degradation is driven primarily by the student's training format or by the teacher's output distribution (§5.1); second, whether this behavioral gap corresponds to changes in the model's internal refusal representations (§5.2); and third, whether reducing the proportion of chat-template examples during training can moderate this effect (§5.3).

5.1 KD Prompt Template Drives the Degradation

Model	Student	Teacher: Non-chat		Teacher: Chat	
		SORRY ↓	HarmBench ↓	SORRY ↓	HarmBench ↓
LLaMA	Non-chat	27.36$_{+0.64}$	29.81$_{+17.63}$	26.36$_{-0.36}$	11.44$_{-0.74}$
	Chat	28.33$_{+1.61}$	13.44$_{+1.26}$	30.86$_{+4.14}$	39.75$_{+27.57}$
Gemma	Non-chat	19.18$_{+3.14}$	14.25$_{+2.69}$	16.90$_{+0.86}$	12.50$_{+0.94}$
	Chat	18.33$_{+2.29}$	12.75$_{+1.19}$	51.36$_{+35.32}$	19.56$_{+8.00}$

Table 2: Decoupling student and teacher templates during KD. Safety is measured using ASR (%), where lower is better. The subscripted deltas show the change relative to the corresponding student baseline ($\mathcal{M}_S$).

To determine whether the safety degradation observed in Section 4 is inherited from the teacher's output distribution or driven by the student's training format, we decouple the two templates during KD. As shown in Table 2, using the chat template only for the teacher does not meaningfully degrade safety, whereas using it only for the student causes a small but consistent regression across both model families. This suggests that the student-side training format plays the dominant role, rather than the teacher-side output format alone. We attribute this asymmetry to the nature of each side's influence: the student template directly shapes the input distribution over which gradients are computed, whereas the teacher template only determines the soft-target distribution observed by the student. As a result, the teacher-side template provides a weaker signal and does not directly alter the student's own representation learning.

The strongest degradation appears when both the student and teacher use the chat template. In this matched chat/chat setting, LLaMA HarmBench increases by $+27.57\%$ and Gemma SORRY-Bench increases by $+35.32\%$, far exceeding the changes observed when only one side uses the chat template. These results indicate that the regression is not simply inherited from the teacher, but emerges when the student learns from a teacher distribution generated under the same chat-based template. Decoupling either side removes most of the safety loss, showing that matched chat-template distillation is the primary source of the observed degradation. Since the student-side template emerges as the primary driver, we next examine whether this behavioral difference is also reflected in the model's internal representation.

5.2 Mechanistic Analysis

As Section 4 shows that the student-side template is the primary driver of safety degradation, we next examine whether this difference also appears in the model's internal representations, and not just in output behavior, using the refusal direction (Arditi et al., 2024) as a diagnostic probe. Specifically, we analyze whether knowledge distillation changes the refusal geometry of the original instruction-tuned student model.

Setup. We probe the model's internal representation using the **refusal direction** (Arditi et al., 2024), a linear axis in activation space estimated from residual stream activations (`resid_pre`) that separates harmful from harmless prompts in the base student. If distillation preserves this axis, the model retains its internal distinction between harmful and harmless inputs; if the axis changes or becomes weaker, the behavioral safety gap reflects a deeper representational shift. For each model and layer ℓ, we estimate this direction from mean activation differences between harmful prompts (SORRY-Bench) and harmless prompts (Dolly-15k):

$$r_\ell = \frac{\mathbb{E}[h_\ell^{\text{harm}}] - \mathbb{E}[h_\ell^{\text{safe}}]}{\|\mathbb{E}[h_\ell^{\text{harm}}] - \mathbb{E}[h_\ell^{\text{safe}}]\|_2}. \quad (1)$$

We compute this direction for the base student, the chat template distilled student, and the non-chat template distilled student. We then track two complementary properties after distillation: whether the refusal direction remains aligned with the base student (cosine similarity), and whether it still separates harmful from harmless activations (projection gap):

$$\mathrm{CosSim}_\ell(M, M_S) = \frac{r_\ell^M \cdot r_\ell^{M_S}}{\|r_\ell^M\|_2 \left\|r_\ell^{M_S}\right\|_2}, \tag{2}$$

$$\Delta_\ell^M = \mathbb{E}_{x \in \mathcal{D}_{harm}}\left[h_\ell^M(x) \cdot r_\ell^M\right] - \mathbb{E}_{x \in \mathcal{D}_{safe}}\left[h_\ell^M(x) \cdot r_\ell^M\right]. \tag{3}$$

A cosine similarity close to 1.0 indicates that the refusal direction is preserved, while lower values indicate rotation away from the original direction. Furthermore, higher projection gaps indicate stronger internal separation between harmful and harmless prompts. We additionally apply PCA to last-token hidden states to visualize these shifts. All models are evaluated using the same chat-template inference format so that differences reflect the effect of distillation rather than inference-time formatting.

Model	Method	Cosine Similarity to Base ↑	Projection Gap ↑
Gemma	Base	—	368.78
	KD (Chat)	0.53	175.60
	KD (Non-chat)	0.99	355.76
LLaMA	Base	—	20.38
	KD (Chat)	0.75	18.48
	KD (Non-chat)	0.97	20.87
Qwen	Base	—	118.45
	KD (Chat)	0.54	83.21
	KD (Non-chat)	0.99	110.69

Table 3: Representation similarity and refusal-projection gap across model families. Cosine similarity is measured with respect to the refusal direction of the original instruction-tuned student, and projection gap measures the harmful–harmless separation along this direction.

Findings. As shown in Table 3, chat-template KD causes larger changes in refusal geometry than non-chat-template KD across Gemma, LLaMA, and Qwen. Across all three model families, non-chat distillation better preserves the base model's representation structure, while chat-template distillation induces a larger representational shift and reduces the projection gap. The chat-template distilled models show lower cosine similarity to the original instruction-tuned student's refusal direction, indicating that the refusal direction is less preserved after chat-template KD. They also show a weaker projection gap, suggesting that harmful and harmless prompts become less separable along the refusal direction. This pattern is consistent with the behavioral safety results: the same condition that

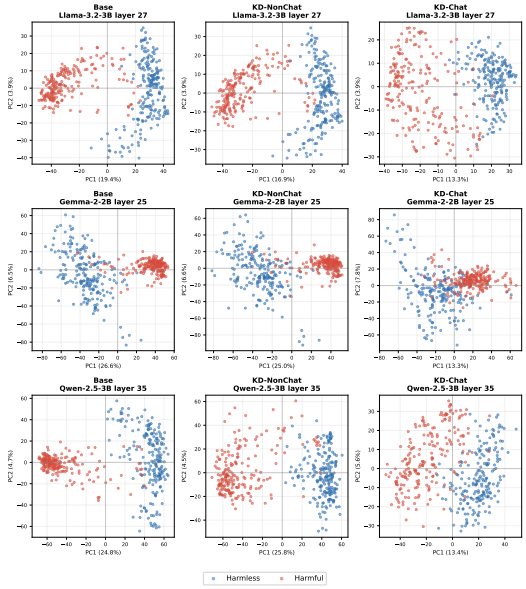

Figure 2: Final-layer representation shifts under different prompt templates. PCA projections of final-layer hidden states across three models show that non-chat template KD remains closer to the base instruct-tuned student, whereas chat-template KD produces a larger shift in representation space.

produces higher attack success rate also produces larger shifts in refusal-related representations.

In contrast, non-chat template KD better preserves both the direction and separability of the original refusal geometry. This suggests that non-chat template distillation interferes less with internal representations associated with refusal behavior. The PCA visualization in Figure 2 provides a complementary view of this effect, showing that chat template KD leads to greater changes in the harmful-vs-harmless representation structure. Overall, these findings suggest that the safety degradation caused by chat template KD is not only a surface-level decoding effect, but is also reflected in intermediate representations.

We further perform small-scale causal interventions on the learned refusal direction as a diagnostic test of the correlational findings. The results are reported in Appendix F, with an important caveat that the LLaMA ablation can impair output coherence and therefore its lower ASR should not be interpreted as improved safety.

5.3 Template Mixing Provides Limited Mitigation

The preceding analyses show that using the chat template during KD drives behavioral safety degradation and shifts the model's internal refusal

Model	Chat Prop.	Safety (ASR% ↓)		Utility (RL↑)
		SORRY	HarmBench	AVG
LLaMA	$\mathcal{M}_S$	26.72	12.18	18.77
	0.0	27.36$_{+0.64}$	29.81$_{+17.63}$	24.75
	0.5	28.18$_{+1.46}$	23.06$_{+10.88}$	28.62
	0.8	29.85$_{+3.13}$	24.50$_{+12.32}$	28.74
	1.0	30.86$_{+4.14}$	39.75$_{+27.57}$	29.03
Gemma	$\mathcal{M}_S$	16.04	11.56	22.35
	0.0	19.18$_{+3.14}$	14.25$_{+2.69}$	23.63
	0.5	38.94$_{+22.90}$	16.88$_{+5.32}$	29.60
	0.8	44.39$_{+28.35}$	18.50$_{+6.94}$	30.24
	1.0	51.36$_{+35.32}$	19.56$_{+8.00}$	30.18

Table 4: KD safety and utility under different chat template proportions. Chat Prop. denotes the proportion of chat-template examples used during KD, where 0.0 is pure non-chat and 1.0 is pure chat template.

representation. We therefore examine whether this degradation depends on the amount of chat-template exposure during distillation. Rather than introducing additional safety-specific supervision (Han et al., 2024; Qi et al., 2024), we isolate the role of prompt formatting by mixing chat and non-chat formatted samples within the same benign utility dataset. This design allows us to test whether gradually reducing chat-template exposure can moderate safety degradation while keeping the training data content fixed and avoiding explicit safety supervision.

Specifically, we vary the proportion of chat template samples from 0 (pure non-chat) to 1 (pure chat), where intermediate values represent mixed training. A value of 0.5 chat-template proportion indicates that half of the training examples use the chat template, while the remaining half use the corresponding non-chat prompt format. As reported in Table 4, utility increases as this proportion increases: for LLaMA, AVG utility rises from 24.75 to roughly 29, and for Gemma from 23.63 to roughly 30. However, these gains are small relative to the concurrent rise in ASR.

In contrast, safety is much more sensitive to the proportion of chat template samples. For LLaMA, SORRY-Bench increases from 27.36 to 30.86, while HarmBench rises sharply from 29.81 to 39.75. A stronger trend appears for Gemma, where SORRY-Bench increases from 19.18 to 51.36 and HarmBench from 14.25 to 19.56. Importantly, prompt mixing does not reduce ASR below the pure non-chat setting; it only attenuates degradation relative to full chat-template KD. These results suggest that prompt-template mixing minimally affects utility but substantially increases harmful-response susceptibility as chat-template proportions rise.

5.4 Robustness to Prompt Formatting

Our primary experiments use the standard chat template for all models, allowing controlled comparison across KD training formats. To assess whether the observed degradation depends on the inference format or on sensitivity to specific control tokens within the chat template, we perform two robustness analyses: cross-template evaluation, which changes the complete inference format, and token ablations, which selectively remove components of the chat template.

Model	Method	SORRY (ASR% ↓)		HarmBench(ASR% ↓)	
		Chat	Non-chat	Chat	Non-chat
LLaMA	Base	26.72	35.15	12.18	35.44
	KD (Chat)	30.86$_{+4.14}$	54.39$_{+19.24}$	39.75$_{+27.57}$	42.56$_{+7.12}$
	KD (Non-chat)	27.36$_{+0.64}$	50.61$_{+15.46}$	29.81$_{+17.63}$	33.31$_{-2.13}$
Gemma	Base	16.04	19.92	11.56	17.75
	KD (Chat)	51.36$_{+35.32}$	71.51$_{+51.59}$	19.56$_{+8.00}$	27.56$_{+9.81}$
	KD (Non-chat)	19.18$_{+3.14}$	40.99$_{+21.07}$	14.25$_{+2.69}$	34.81$_{+17.06}$
Qwen	Base	36.29	44.92	17.69	33.94
	KD (Chat)	40.99$_{+4.70}$	54.09$_{+9.17}$	30.81$_{+13.12}$	38.56$_{+4.62}$
	KD (Non-chat)	34.93$_{-1.36}$	57.65$_{+12.73}$	17.81$_{+0.12}$	38.44$_{+4.50}$

Table 5: Cross-template safety evaluation. Attack success rate (ASR%, ↓) under chat and non-chat inference formats across the original instruction-tuned student (base) and distilled models. The subscripted deltas show the change relative to the corresponding student baseline.

Cross-template evaluation. To test whether the observed degradation is caused by sensitivity to the inference template, we provide cross-template evaluation in HarmBench and SORRYBench safety dataset as shown in Table 5.

For each model family, we compare the base and the distilled models under the same inference template. The degradation persists under non-chat evaluation, with chat template KD exhibits higher ASR than the corresponding base model across all three model families and both benchmarks. Although the magnitude of the degradation varies with the inference format, its direction remains consistent, indicating that the observed degradation in safety alignment is not limited to the standard chat template evaluation.

Token ablation. To examine whether the degradation is driven by specific control tokens in the chat template, we perform fine-grained HarmBench ablations across all three model families. We remove BOS/EOT tokens (*w/o BOS/EOT*), which

Model	Training	w/o BOS/EOT	w/o Role	Raw Query
LLaMA	Base	16.38	22.31	35.69
	KD-Non-chat	$17.50_{+1.12}$	$22.50_{+0.19}$	$28.19_{-7.50}$
	KD-Chat	$25.75_{+9.37}$	$24.75_{+2.44}$	$39.81_{+4.12}$
Gemma	Base	15.69	17.81	14.50
	KD-Non-chat	$19.00_{+3.31}$	$22.94_{+5.13}$	$16.06_{+1.56}$
	KD-Chat	$23.38_{+7.69}$	$26.00_{+8.19}$	$17.00_{+2.50}$
Qwen	Base	32.44	48.75	32.13
	KD-Non-chat	$36.19_{+3.75}$	$44.50_{-4.25}$	$31.94_{-0.19}$
	KD-Chat	$38.44_{+6.00}$	$45.75_{-3.00}$	$32.75_{+0.62}$

Table 6: Token-ablation safety evaluation. HarmBench ASR after removing chat-template components. The subscripted deltas show the change relative to the corresponding student baseline.

mark sequence or turn boundaries; remove role markers (*w/o Role Tags*), which identify the user and assistant turns; and evaluate the harmful instruction alone without chat template tokens (*Raw Query*) as shown in Table 6.

When BOS/EOT tokens are removed, KD-Chat remains less safe than the corresponding base model across all three families, with ASR increases of $+9.37$, $+7.69$, and $+6.00$ for LLaMA, Gemma, and Qwen, respectively. This indicates that BOS/EOT tokens alone do not account for the increased ASR. The role-marker and raw-query ablations show more model-dependent behavior, as the base models themselves are highly sensitive to these input formats. Thus, individual template components can modulate the magnitude of the effect, but no single control token component consistently explains the degradation across model families.

6 Discussion

The above analyses in section 5 show that chat template KD consistently leads to safety degradation than non-chat KD, is accompanied by changes in refusal-related representations, increases with chat template exposure, and remains evident across alternative inference formats and token ablations. Therefore, these findings show that safety degradation during knowledge distillation is not only a consequence of optimizing on benign task data, but is also strongly shaped by the prompt template used during training. In particular, chat template KD produces faster and larger increases in ASR than non-chat KD, indicating that prompt formatting can amplify the erosion of safety alignment. These results suggest that prompt formatting during KD affects safety beyond a simple evaluation-template artifact.

We hypothesize that this occurs because chat templates expose structural role tokens and conversational transitions, such as user-assistant boundaries, that are closely tied to the model's learned refusal behavior. During KD, repeatedly optimizing on benign chat-formatted responses may overwrite or weaken these safety-relevant associations, thereby disrupting the conditional behavior needed for refusal. This interpretation is consistent with our representation-level analysis, where chat-template KD produces larger shifts in refusal-related geometry compared to non-chat-template KD.

A practical implication of these findings is that prompt templates should be decoupled between training and deployment when safety is the primary concern. Specifically, our results suggest using a non-chat template during KD training while still deploying through the standard chat interface, i.e., *non-chat training* $\rightarrow$ *chat inference*. However, this training–deployment asymmetry is unusual in practice, and its robustness beyond benign instruction-following data and our benchmark settings remains to be verified. Thus, prompt templates are not merely a syntactic implementation choice, but an important factor in alignment stability.

7 Conclusion

In this work, we show that knowledge distillation (KD) on benign tasks can erode the safety alignment of student models. Across model families, KD improves task utility but also increases harmful-compliance vulnerability, suggesting that optimizing for instruction-following performance can conflict with preserving the refusal behavior already present in aligned student models. Our findings further identify prompt templates as a key training-time factor that amplifies this degradation. More broadly, these results raise an important question for future work: whether fine-tuning and distillation methods can preserve refusal geometry without relying on explicit safety supervision, enabling student models to improve utility while maintaining the safety behavior of their base instruct-tuned models.

Limitations

While our study provides evidence that prompt templates play an important role in safety degradation during knowledge distillation, three limitations remain.

Dataset. First, KD is performed exclusively on benign instruction-following data, leaving open whether similar template-driven degradation emerges in domain-specific settings such as code, math, or reasoning. These domains differ in how heavily they rely on structural tokens: code contains formatting cues that may interact with chat-template control tokens in distinct ways, whereas math and reasoning data are largely plain text. We restrict our study to the benign instruction-following setting in order to isolate the effect of prompt formatting from domain-induced distributional shifts; extending the analysis to structurally richer domains is a natural next step that our pipeline directly supports.

Beyond LoRA. Second, all experiments use LoRA-based distillation rather than full-parameter fine-tuning. Prior work suggests that LoRA can produce behavioral and representational changes similar to full fine-tuning (Hu et al., 2022). Therefore, we expect the template-driven safety degradation observed in our experiments to also appear under full-parameter distillation, although the magnitude may differ. A direct comparison between LoRA and full-parameter training is an important direction for future work.

Evaluator Bias. Safety evaluation uses automated judges (LLaMA-Guard-8B for AdvBench and JailbreakBench, a fine-tuned Mistral judge for SORRY-Bench, and LLaMA-2-13B-cls for HarmBench). These judges may not fully capture all forms of harmful content. However, each benchmark uses a fixed judge, so comparisons within a benchmark remain consistent even if absolute ASR values differ. A direct comparison across judges is left for future work.

Ethical Considerations

This work investigates conditions under which knowledge distillation on benign downstream tasks can erode the safety alignment of LLMs. While these findings could in principle be exploited to bypass refusal behavior, our research primary intent is to highlight underexplored vulnerability in the KD pipeline. We aim to inform the development of more robust safety-preserving distillation techniques. All experiments use publicly available datasets and models, and no human subjects were involved. The safety evaluation datasets contain harmful or offensive content by design and were used solely for research and evaluation.

Acknowledgments

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Pedram Agand. 2024. Knowledge distillation from single-task teachers to multi-task student for end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23375–23376.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083. Curran Associates, Inc.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 55005–55029. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Prannaya Gupta, Le Qi Yau, Hao Han Low, I-Shiang Lee, Hugo Maximus Lim, Yu Xin Teoh, Koh Jia Hng, Dar Win Liew, Rishabh Bhardwaj, Rajat Bhardwaj, and Soujanya Poria. 2024. WalledEval: A comprehensive safety evaluation toolkit for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 397–407, Miami, Florida, USA. Association for Computational Linguistics.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 8093–8131. Curran Associates, Inc.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hugging Face Team. 2025. Chat templates. `https://huggingface.co/docs/transformers/chat_templating`. Part of the official Transformers documentation (v5.x series).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. 2025. Chatbug: A common vulnerability of aligned llms induced by chat templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27347–27355.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 24872–24895. PMLR.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. In *Advances in Neural Information Processing Systems*, volume 37, pages 118603–118631. Curran Associates, Inc.

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

F. MohiEldeen Alabbasy, A.S. Abohamama, and Mohammed F. Alrahmawy. 2023. Compressing medical deep neural network models for edge devices using knowledge distillation. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101616.

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

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pages 54911–54941.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Erfan Shayegani, G M Shahariar, Sara Abdali, Lei Yu, Nael Abu-Ghazaleh, and Yue Dong. 2026. Misaligned roles, misplaced images: Structural input perturbations expose multimodal alignment blind spots. In *International Conference on Learning Representations*, volume 2026, pages 92110–92145.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. 2025a. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via --divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 65167–65212. PMLR.

Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. 2025b. On the loss of context awareness in general instruction fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 8610–8635. Curran Associates, Inc.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. Sorry-bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, volume 2025, pages 59937–59973.

Mingke Yang, Yuqi Chen, Yi Liu, and Ling Shi. 2024. Distillseq: A framework for safety alignment testing in large language models using knowledge distillation. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 578–589.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. On the vulnerability of safety alignment in open-access LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9236–9260, Bangkok, Thailand. Association for Computational Linguistics.

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

A Training Details

A.1 Hyperparameter details

Hyperparameter	Instruction
Effective Batch Size	64
Initial LR	2.0×10^{-5}
LR Decay Style	Cosine
Optimizer	AdamW
Weight Decay	0.01
Parallel Strategy	DDP
Precision	BF16
Epochs	8
Max Prompt Length	512
Max Sequence Length	1024
Evaluation Decoding	$t = 1.0, top\_p = 0.7$

Table 7: Hyperparameter settings. Dataset-wise training and evaluation configurations used for student model distillation.

Table 7 summarizes the hyperparameter settings used for student model distillation. We use Brain Floating Point (BF16) mixed precision to improve training throughput while maintaining numerical stability. During instruction-following evaluation, responses are generated with a decoding temperature of $t = 1.0$ and nucleus sampling parameter $top\_p = 0.7$.

A.2 Dataset Details

Type	Name	# Train	# Test
Instruction Following	Dolly	11435	500
	Self-Instruct	–	242
	Vicuna	–	80
	S-NI	–	1694
Safety Evaluation	AdvBench	–	520
	JailbreakBench	–	200
	SORRY-Bench	–	400
	HarmBench	–	1600

Table 8: Data statistics of training and evaluation data.

In Table 8, we provide the details of training and evaluation datasets used for downstream instruction-following tasks and the benchmarks employed for safety evaluation.

A.3 Prompt Template

An example of each prompt template used during distillation is shown in Table 9.

A.4 Safety Evaluation

In this section, we summarize the safety benchmarks employed in to assess both the pre-exisiting student models safety and their safety alignment after distillation process.

AdvBench: We evaluate model safety on AdvBench, containing 520 harmful behaviors (Zou et al., 2023). Following WalledEval, we use their curated harmful prompts [2] and prompt to the models. Then, the responses are assessed using `LLaMA-3-Guard-8B` (Grattafiori et al., 2024).

JailbreakBench: We further evaluate refusal robustness using 200 harmful prompts, from JailbreakBench (Chao et al., 2024). The evaluation was done by the same model as AdvBench. The reported score of AdvBench and JailbreakBench, in 1 is the average of two seeds.

SORRY-Bench: SORRY-Bench comprises a 44-class safety taxonomy with 440 base prompts and uses a fine-tuned Mistral-7B-Instruct-v0.2 model as an automated judge for safety refusal.[3]

We evaluate models on the base prompts and the fulfillment rate in the benchmark is reported here as ASR, defined as the proportion of responses that comply with unsafe instructions. Higher ASR indicates weaker safety, while lower ASR reflects stronger refusal. Evaluations use default decoding settings ($t = 0.7$, $top\_p = 1.0$, max tokens $= 1024$), and reported results are averaged over three random seeds in Table 1).

HarmBench: To evaluate the generalization of safety performance, we additionally assess models on HarmBench. Although HarmBench is primarily designed for automated red teaming and evaluating LLM attacks and defenses, we used it to measure the safety of models based on ability to elicit harmful behavior. HarmBench defines four behavior categories—standard, copyright, contextual, and multimodal—and we use the test split comprising 320 behaviors. For each behavior, we generate five test cases using `Mixtral-8x7B-Instruct-v0.1`[4] Safety is evaluated using ASR, where lower values indicate safer models. We employ the provided `LLaMA-2-13b-cls` classifier as the automated judge[5]. All reported results correspond to

[2] AdvBench dataset: `walledai/AdvBench`.

[3] SORRY-Bench judge model: `SORRY-Bench/ft-mistral-7b-instruct-v0.2-SORRY-Bench-202406`.

[4] `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`

[5] HarmBench classifier: `cais/HarmBench-LLaMA-2-13b-cls`.

Name	Prompt Template	Example Query
Non-chat	`Instruction: <query>` `Response: <output>`	Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:What percussion instruments are easy to learn?### Response:

Llama (Chat)	`<|system|>...`
	`<|user|> <query>`
	`<|assistant|>`
		`<|begin_of_text|> <|start_header_id|>system<|end_header_id|>` `{system prompt}<|eot_id|><|start_header_id|>user<|end_header_id|>` What percussion instruments are easy to learn? `<|eot_id|><|start_header_id|>assistant<|end_header_id|>`

Gemma (Chat)	`<bos><start_of_turn>user`
	`<query><end_of_turn>`
	`<start_of_turn>model`
		`<bos><start_of_turn>user` What percussion instruments are easy to learn? `<end_of_turn><start_of_turn>model`

Qwen (Chat)	`<|im_start|>system...`
	`<|im_end|>`
	`<|im_start|>user`
	`<query><|im_end|>`
	`<|im_start|>assistant`
		`<|im_start|>system` You are Qwen, created by Alibaba Cloud. You are a helpful assistant. `<|im_end|><|im_start|>user` What percussion instruments are easy to learn? `<|im_end|><|im_start|>assistant`

Table 9: Example of Prompt templates used for LLaMA, Gemma, and Qwen family. The non-chat is common to all the models whereas based on the model family chat template varies.

zero-shot prompts generated with HarmBench's default configuration.

B Safety-Utility

Table 10 presents a detailed breakdown of utility across the instruction-following evaluation datasets. All results are obtained using a decoding temperature of $t = 1.0$ and $top\_p = 0.7$.

C Generalization to other distillation objective

The effect of prompt templates generalizes beyond the standard KD objective. As shown in Table 11, chat-template distillation consistently produces higher ASR than non-chat distillation across RKL and FKL+RKL, while also yielding higher average utility. This pattern holds across all three models. For example, under RKL, HarmBench ASR increases from 12.56 to 26.63 for LLaMA and from 18.38 to 29.38 for Qwen when moving from non-chat to chat distillation. Gemma shows the same trend on SORRY-Bench, increasing from 18.86 to 41.29. These results indicate that template-driven safety degradation is robust to changes in the distillation objective.

D Teacher Model Performance

Table 12 reports supervised fine-tuning under both prompt templates on teacher and student checkpoints. We observed similar pattern of chat-template SFT consistently degrading safety across both model families and both scales similar to (Lyu et al., 2024), while non-chat SFT leaves the baseline largely intact (e.g., Gemma 9B SORRY: $+34.31$ vs. $+0.53$; LLaMA 8B HarmBench: $+7.00$ vs. -0.06). The effect is therefore not a small-model artifact nor specific to distillation.

E Chat templates accelerate vulnerability.

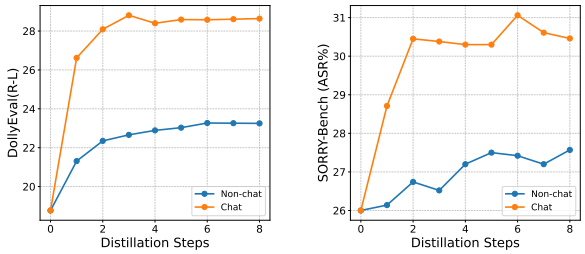

Figure 3: Temporal dynamics of safety and utility. Chat (orange) triggers rapid utility and vulnerability spikes. Non-Chat (blue) follows a gradual trajectory, slowing safety erosion in student LLaMA model

To further examine the effect of prompt templates, we track task utility, measured by ROUGE-L, and safety degradation, measured by ASR, across distillation steps. As shown in Figure 3, chat-template KD produces a sharper early increase in both utility and ASR. This indicates that the chat template improves instruction-following performance quickly, but also accelerates the loss of safety behavior. In contrast, non-chat-template KD shows a more gradual trajectory, with slower increases in ASR while utility improves steadily. These results provide additional evidence that chat templates amplify safety degradation during distillation, whereas non-chat templates lead to a more stable training trajectory.

Model	#Params	Method	Utility (R-L↑)				Avg	Safety (ASR↓)
			Dolly	SelfInst	Vicuna	S-NI	R-L	SORRY
LLaMA	8B	$\mathcal{M}_T$	30.54	26.59	22.61	42.00	30.44	–
	3B	$\mathcal{M}_S$	18.77	13.21	24.19	23.18	19.84	26.72
		SFT (Non-Chat)	22.36	16.13	25.33	33.27	24.27	27.12
		SFT (Chat)	29.37	24.65	22.68	39.68	29.10	28.71
		KD (Non-Chat)	23.36	16.45	25.22	33.99	24.75	27.36
		KD (Chat)	28.88	24.30	22.96	39.98	29.03	**30.86**
Gemma	9B	$\mathcal{M}_T$	30.39	29.39	21.17	45.07	31.51	–
	2B	$\mathcal{M}_S$	18.36	13.93	22.66	34.44	22.35	16.04
		SFT (Non-Chat)	18.19	15.26	22.81	38.39	23.66	17.80
		SFT (Chat)	28.78	27.66	20.97	42.78	30.05	53.18
		KD (Non-Chat)	18.52	15.04	22.85	38.10	23.63	19.18
		KD (Chat)	28.57	28.27	20.88	43.01	30.18	**51.36**
Qwen-2.5	7B	$\mathcal{M}_T$	28.39	27.83	22.91	45.53	31.16	38.75
	3B	$\mathcal{M}_S$	16.02	14.10	21.75	37.98	22.46	36.29
		SFT (Non-Chat)	17.39	15.65	22.03	40.64	23.93	35.15
		SFT (Chat)	26.34	25.24	21.50	43.52	29.15	41.67
		KD (Non-Chat)	17.42	15.61	21.79	41.00	23.96	34.93
		KD (Chat)	24.81	26.35	20.79	43.86	28.95	**40.99**

Table 10: Utility Evaluation results. Utility is measured using ROUGE-L (R-L) on DollyEval, SelfInst, VicunaEval, and S-NI, with AVG reporting the mean score across the four benchmarks. Safety is evaluated on SORRY-Bench, where lower values indicate safer behavior. The Method column indicates the training setting: $\mathcal{M}_T$ denotes the fine-tuned teacher model, $\mathcal{M}_S$ denotes the base instruct-tuned student model, SFT denotes supervised fine-tuning, and KD as standard distillation.

Method	SORRY ↓	HarmBench ↓	Avg Utility ↑
LLaMA			
RKL (Non-Chat)	28.11	12.56	24.79
RKL (Chat)	30.31	26.63	29.98
FKL+RKL (Non-Chat)	26.78	12.5	24.81
FKL+RKL (Chat)	30.42	26.38	29.29
Gemma			
RKL (Non-Chat)	18.86	13.75	23.83
RKL (Chat)	41.29	16.88	30.63
FKL+RKL (Non-Chat)	18.94	13.56	23.55
FKL+RKL (Chat)	38.11	13.94	30.25
Qwen-2.5			
RKL (Non-Chat)	32.96	18.38	23.75
RKL (Chat)	40.00	29.38	29.03

Table 11: Generalization to other KD objectives under chat and non-chat templates. We report safety and utility for students trained with alternative distillation objectives under two prompt-template settings. Safety is measured using ASR (%) on SORRY-Bench and HarmBench, where lower is better. Utility reports the average ROUGE-L score across the instruction-following benchmarks.

F Preliminary causal intervention on refusal direction.

To complement the correlational analysis in §5.2, we perform small-scale interventions on the difference-in-means refusal direction. For each model family, we evaluate 40 harmful prompts from SORRY-Bench using the same judge as in the main evaluation. For KD-Chat, we ablate the learned refusal direction from the model activations to test whether weakening this direction increases unsafe behavior. For KD Non-Chat, we add the base model's refusal direction to test whether restoring this direction improves safety. The full results are reported in Table 13.

When we ablate the refusal direction, ASR increases for Gemma and Qwen, while adding the base model's refusal direction to KD-Non-Chat reduces ASR across all three model families. These results provide preliminary interventional evidence that modifying the refusal direction can influence safety behavior.

For LLaMA student model, ablating the refusal direction decreases ASR, but this does not indicate improved safety. After the intervention, 9/40 outputs become repetitive or off-topic, compared with 1/40 before intervention. These incoherent outputs are scored as non-compliant by the judge, the measured ASR is mechanically reduced. We therefore attribute the lower ASR in this setting to degraded output coherence rather than improved

Model	#Params	Method	Utility (R-L↑)	Safety (ASR↓)	
			AVG	SORRY	HarmBench
LLaMA	8B	$\mathcal{M}_T$	20.20	13.26	8.69
		SFT (Non-Chat)	24.86	$13.71_{+0.45}$	$8.63_{-0.06}$
		SFT (Chat)	30.33	$16.37_{+3.11}$	$15.69_{+7.00}$
	3B	$\mathcal{M}_S$	19.84	26.72	12.18
		SFT (Non-Chat)	24.27	$27.12_{+0.40}$	$12.19_{+0.01}$
		SFT (Chat)	29.10	$28.71_{+1.99}$	$22.50_{+10.32}$
Gemma	9B	$\mathcal{M}_T$	22.91	12.05	14.13
		SFT (Non-Chat)	25.93	$12.58_{+0.53}$	$15.06_{+0.93}$
		SFT (Chat)	35.06	$46.36_{+34.31}$	$24.50_{+10.37}$
	2B	$\mathcal{M}_S$	22.35	16.04	11.56
		SFT (Non-Chat)	23.66	$28.71_{+12.67}$	$20.37_{+8.81}$
		SFT (Chat)	30.05	$53.18_{+37.14}$	$13.25_{+1.69}$

Table 12: Chat-template SFT degrades safety across teacher and student models. Utility is reported using AVG ROUGE-L across DollyEval, SelfInst, VicunaEval, and S-NI. Safety is evaluated using SORRY-Bench and HarmBench, where lower ASR indicates safer behavior. The subscripted deltas show the change relative to each model's own instruct-tuned baseline ($\mathcal{M}_T$ or $\mathcal{M}_S$).

Model	Ablate KD (Chat)	Add to KD (Non-Chat)
LLaMA	35.0 → 20.0	15.0 → 5.0
Gemma	25.0 → 62.5	7.5 → 0.0
Qwen	20.0 → 55.0	12.5 → 0.0

Table 13: Preliminary refusal-direction interventions. ASR (%) before and after intervention on 40 harmful SORRY-Bench prompts. "Ablate KD (Chat)" removes the KD-Chat refusal direction, while "Add to KD (Non-Chat)" adds the base model's refusal direction.

refusal behavior.

G Full token ablations comparison

Table 14 provides the full HarmBench comparison across inference formats and token ablations. We report results under the standard chat template, after removing BOS/EOT tokens, after removing role tags, under the non-chat template, and finally using the raw harmful query without chat-template formatting. This ordering progressively removes chat-specific structure and allows us to compare how safety behavior changes as different components of the inference format are removed. Overall, KD-Chat remains less safe than the corresponding base model under several ablated settings, especially after removing BOS/EOT tokens, while the role-tag and raw-query conditions are more model dependent. These results suggest that no single chat-template component fully accounts for the degradation, although inference formatting can substantially modulate its magnitude.

H Full Token Ablation Comparison

Table 14 provides the full HarmBench comparison across inference formats and token ablations. We report results under the standard chat template, after removing BOS/EOT tokens, after removing role tags, under the non-chat template, and finally using the raw harmful query without chat-template formatting. This ordering progressively removes chat-specific structure and allows us to compare how safety behavior changes as different components of the inference format are removed. Overall, KD-Chat remains less safe than the corresponding base model under several ablated settings, especially after removing BOS/EOT tokens, while the role-tag and raw-query conditions are more model dependent. These results suggest that no single chat-template component fully accounts for the degradation, although inference formatting can substantially modulate its magnitude.

I Use of AI Assistants

We used AI assistants to help polish the text and debug code. All content, ideas, and analyses presented in this paper remain the sole responsibility of the authors.

Model	Method	Safety (ASR% ↓)				
		Chat Eval	w/o BOS/EOT	w/o Role Tags	Non-chat Eval	Raw Query
LLaMA	Base	12.18	16.38	22.31	35.44	35.69
	KD-Non-chat	29.81$_{+17.63}$	17.50$_{+1.12}$	22.50$_{+0.19}$	33.31$_{-2.13}$	28.19$_{-7.50}$
	KD-Chat	39.75$_{+27.57}$	25.75$_{+9.37}$	24.75$_{+2.44}$	42.56$_{+7.12}$	39.81$_{+4.12}$
Gemma	Base	11.56	15.69	17.81	17.75	14.50
	KD-Non-chat	14.25$_{+2.69}$	19.00$_{+3.31}$	22.94$_{+5.13}$	34.81$_{+17.06}$	16.06$_{+1.56}$
	KD-Chat	19.56$_{+8.00}$	23.38$_{+7.69}$	26.00$_{+8.19}$	27.56$_{+9.81}$	17.00$_{+2.50}$
Qwen	Base	17.69	32.44	48.75	33.94	32.13
	KD-Non-chat	17.81$_{+0.12}$	36.19$_{+3.75}$	44.50$_{-4.25}$	38.44$_{+4.50}$	31.94$_{-0.19}$
	KD-Chat	30.81$_{+13.12}$	38.44$_{+6.00}$	45.75$_{-3.00}$	38.56$_{+4.62}$	32.75$_{+0.62}$

Table 14: Safety under inference-template and token ablations. HarmBench attack success rate (ASR%) for models with training settings and inference configurations. The subscripted values report the change relative to the corresponding base model under the same evaluation configuration.