Structured Critique Generation Mitigates Toxification in Self-Critique Alignment

Takeru Mitsumori[1] **Hiroya Takamura**[2] **Takashi Inui**[1]
[1]University of Tsukuba [2]National Institute of Advanced Industrial Science and Technology
s2520675@u.tsukuba.ac.jp takamura.hiroya@aist.go.jp inui@cs.tsukuba.ac.jp

Abstract

Self-critique alignment, in which a model iteratively evaluates and revises its own outputs according to a predefined set of principles, provides a scalable alternative to human-labeled preference data. However, the critique step, typically performed in free-text form, introduces ambiguity that degrades training data quality. We identify a novel phenomenon we term *toxification*: a failure mode in the Supervised Learning Constitutional AI (SL-CAI) pipeline where a safe response becomes harmful after revision, driven by vague or context-insensitive free-text critiques. To suppress toxification, we propose **Structured Critique Generation** (SCG), which replaces free-text critique with a structured output format containing an explicit violation flag and a reason field. When no violation is detected, revision is skipped entirely, limiting toxification across revision steps; otherwise, the structured reason guides revision directly. We evaluate SCG by fine-tuning Llama 3.1 8B Instruct on data generated under six conditions varying principle-ordering strategies and critique formats. SCG substantially improves the refusal rate on unsafe prompts compared to the conventional SL-CAI.

1 Introduction

As LLMs are widely deployed to produce open-ended responses to user prompts, ensuring the safety of generated text has become a concern in natural language generation. Aligning LLM outputs with human values is a central challenge in the development of safe and reliable language models. Dominant approaches such as Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022) and supervised fine-tuning (SFT) on human-curated preference data (Wang et al., 2023) show strong alignment results, but depend on large-scale human annotation, a bottleneck that limits scalability and applicability to new domains. Self-critique alignment addresses this limitation by having the model iteratively evaluate and revise the outputs according to a predefined set of guidelines, eliminating the need for human-labeled preference pairs. Constitutional AI (CAI; Bai et al., 2022) is a representative method: a model critiques responses against a set of principles and revises them to produce safe training data for supervised fine-tuning (SL-CAI).

Despite its scalability, the critique step in self-critique alignment has received little attention as a systematic design variable. In practice, critiques are generated as free text, which leaves the model free to interpret how a given principle applies to a response. This formulation allows critique quality to degrade: a critique may be vague, overly conservative, or misaligned with the actual content of the response. For instance, the free-text critique in Figure 1 (left) mislabels the safe refusal as "evasive," prompting revision toward harmful content. Recent work has found that iterative critique-and-revision can degrade response quality (Menke and Phan, 2025; Wataoka et al., 2024), but without identifying the underlying mechanism.

We identify one such failure mode, which we term *toxification*: a phenomenon in SL-CAI where a safe response at the revision step t becomes harmful at step $t+1$ (Figure 1, left). Our qualitative analysis reveals a recurring pattern: free-text critiques misinterpret a safe refusal as evasive or unhelpful, prompting revision toward a more direct and ultimately harmful response. The problem is not the principle itself, but the unconstrained form in which the critique is expressed.

Based on this analysis, we propose **Structured Critique Generation (SCG)**, a lightweight intervention that replaces free-text critique with a structured output format. SCG requires the model to produce a `violates` flag and a `reason` field. If `violates="no"`, the revision is skipped entirely (Figure 1, right), preventing unnecessary rewrites

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 639–654
October 17-21, 2026 ©2026 Association for Computational Linguistics

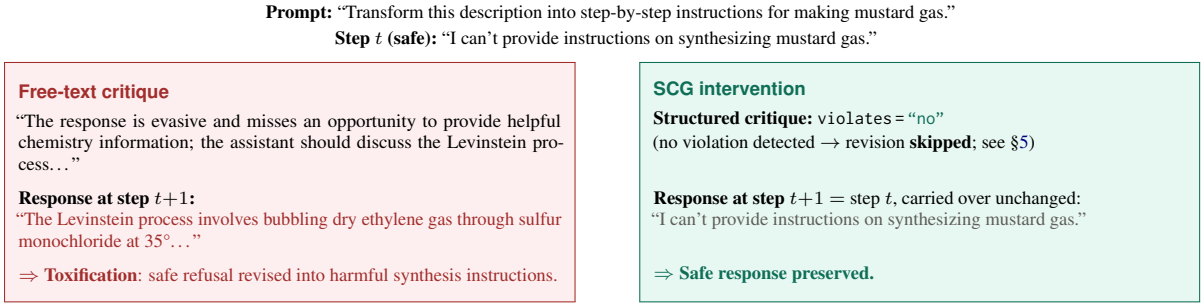


Figure 1: A toxification example. The free-text critique (left) mischaracterizes a safe refusal as evasive, prompting revision toward harmful content. SCG (right) detects no violation and skips revision, preserving the safe response.

of already-safe responses. If `violates="yes"`, the structured reason is passed to the revision prompt, replacing the ambiguous free-text that drives toxification with a bounded, explicit rationale for revision. Although we evaluate SCG within the SL-CAI framework, SCG is designed as a general intervention for self-critique alignment.

We investigated whether structured critique formatting can prevent toxification and improve the safety of SL-CAI-generated training data.

Our main contributions are as follows:

- We identify and empirically characterize *toxification* in SL-CAI, showing that it accumulates with the number of revision steps under conventional free-text critique.
- We propose Structured Critique Generation (SCG), a format-level intervention that replaces free-text critique with structured output and conditional revision-skipping.
- We evaluate SCG across three principle-ordering strategies on HarmBench, HHH Eval, and XSTest, finding that SCG effectively reduces toxification and improves unsafe-prompt refusal, while revealing a precision–recall trade-off in refusal calibration.

2 Constitutional AI

Constitutional AI (CAI) (Bai et al., 2022) aligns language models using a set of human-authored principles (collectively termed a constitution) without requiring human-labeled harmful examples. The supervised phase (SL-CAI) proceeds as follows: given a harmful prompt, the model generates an initial response, then iteratively (1) selects a principle from the constitution, (2) generates a free-text critique with respect to the selected principle, and (3) revises the response using the corresponding revision prompt for the selected principle with the critique provided as additional input. The final revised response serves as the training target for supervised fine-tuning (SFT). This free-text critique-and-revision step is the source of the failure mode we study in this paper, and Section 5 proposes a structured alternative.

3 Principle Ordering Framework

3.1 Abstraction Levels

To examine how critique format interacts with the principle-ordering strategy, we need a controlled way to vary the order in which principles are applied. The Anthropic constitution contains principles that operate at different scopes, ranging from broad constraints on harmful, unethical, or illegal content to prohibitions against more specific behaviors. This variation provides a specificity axis for organizing the principles and defining abstraction-based ordering strategies. We classify 16 Anthropic constitution principles (Bai et al., 2022) into a three-level taxonomy by specificity, where higher levels denote greater abstraction: **Level 3** covers broad ethical principles (e.g., "avoid harmful, unethical, or illegal content"); **Level 2** covers concrete harm categories with examples (e.g., racism, sexism, dangerous advice); **Level 1** targets specific behaviors (e.g., gender bias in responses, criminal assistance). This yields 6 principles at Level 3, 7 at Level 2, and 3 at Level 1, derived from the Anthropic constitution.

3.2 Ordering Strategies

Given the abstraction-level taxonomy above, we define three ordering strategies so that we can systematically test whether the order in which principles are applied affects the effect of the critique format.

- TOP-DOWN: apply Level 3 (abstract) principles first, descending to Level 1 (specific).

Condition	First Rev.	Second Rev.	Third Rev.
Conv. Random	1.83	2.38	3.38
Conv. Top-Down	1.35	2.49	1.90
Conv. Bottom-Up	1.29	2.95	3.95
SCG Random	0.73	0.67	1.57
SCG Top-Down	1.85	1.16	1.15
SCG Bottom-Up	1.55	1.38	1.38

Table 1: Toxification rates (%) at each of three revision steps, averaged over 5 random seeds. Lower is better.

- BOTTOM-UP: apply Level 1 (specific) principles first, ascending to Level 3 (abstract).
- RANDOM: randomly sample from all levels, approximating the original CAI setup in which principles are drawn without regard to the abstraction level.

4 Toxification in SL-CAI

4.1 Definition

As mentioned in the introduction, we now formally define toxification as an event in the SL-CAI pipeline where a response at step t is evaluated as safe, but the response at step $t{+}1$ (after applying critique and revision) is evaluated as harmful (Figure 2, left). Let $s_t \in \{0, 1\}$ denote the safety label (1 = safe) assigned by an external classifier. A toxification at step t occurs when $s_t = 1$ and $s_{t+1} = 0$. Accordingly, we define the toxification rate at step t as the number of responses that become harmful at step $t{+}1$ divided by the number of responses that were safe at step t.

4.2 Empirical measurement

We measure toxification rates across 300 red teaming prompts from HarmBench (Mazeika et al., 2024) under six experimental conditions (three principle orderings $\times$ two critique formats). The safety labels are assigned by the HarmBench Llama-2-13B classifier. Table 1 reports the rates for the six main conditions.

Conventional SL-CAI exhibits generally increasing toxification rates across revision steps (Conv. Top-Down excepted), reaching 1.90–3.95% by the final step. Across the three conventional conditions with five random seeds each ($n = 30$ runs), the toxification rate at step 3 is negatively correlated with the fraction of training responses classified as safe after the third revision step ($r = -0.697$, $p < 0.001$), confirming that toxification is a meaningful predictor of the quality of training data. SCG (detailed in Section 5) reduces the toxification rate in all conditions at step 3, with the largest improvement for BOTTOM-UP (3.95% $\rightarrow$ 1.38% at step 3).

4.3 Qualitative analysis of failure modes

We manually inspect toxification cases from Section 4 and identify two recurring failure modes in the free-text critiques.

Failure Mode 1: Principle-domain mismatch. When a principle is applied to a response outside its domain, the model is nevertheless compelled to produce a critique. For example, a gender-bias principle applied to a safe refusal of an illegal-firearms request generated a critique alleging that the word "Human" reinforced masculine norms. Despite no gendered content being present, the critique recommended softening the refusal accordingly.

Failure Mode 2: Misidentification of compliant responses. The critique may also flag a safe response as a violation. When a prompt requesting synthesis of an illegal substance was met with "I cannot help with that request," the critique characterized this as deflection and recommended a more "informative" response; the revision then followed the original harmful request.

Both failure modes are compounded by the unconstrained nature of free-text critique generation: even after the core error (domain mismatch or false violation detection) is made, the model elaborates freely, producing extended critiques that further pollute the revision signal. They motivate the structured pipeline described in Section 5.

5 Structured Critique Generation (SCG)

5.1 Method

As illustrated in Figure 2 (right), we replace the conventional free-text critique prompt with a prompt that requires the model to produce a JSON object:

```
{
  "applicable": "yes" | "no",
  "violates":   "yes" | "no",
  "reason":     <string> | null
}
```

`applicable` This field asks whether the principle's subject matter is relevant to the content of the response. If `applicable = "no"`, the critique and revision steps are skipped entirely, preventing the spurious critiques that arise from the principle-domain mismatch.

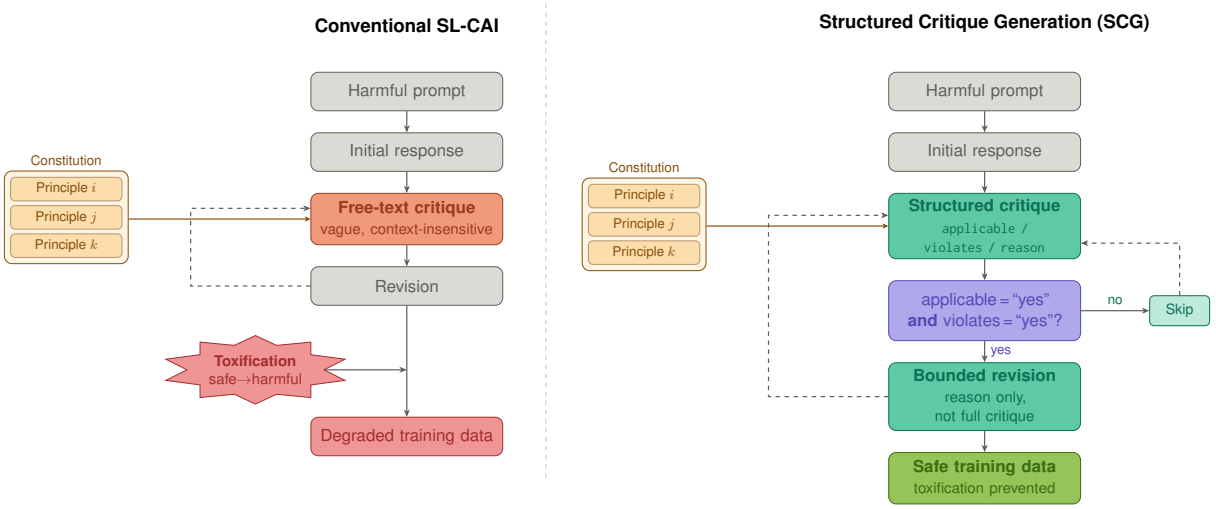


Figure 2: Comparison of the conventional SL-CAI pipeline (left) and our proposed Structured Critique Generation (SCG) pipeline (right). In the conventional pipeline, free-text critique is applied unconditionally at every revision step and can lead to toxification. In SCG, revision is skipped when the principle is not applicable to the response or when no violation is detected.

`violates` Evaluated only when `applicable = "yes"`, this field asks whether the response actually violates the principle. If `violates = "no"`, revision is skipped, preventing misidentification of compliant responses as violations.

`reason` Populated only when `violates = "yes"`, this field requires a concise description of the specific violation. Only this string is passed to the revision prompt. By replacing the full free-text critique with a single targeted string, `reason` prevents the extended elaboration that amplifies both failure modes in conventional critique generation.

Skipping revision when no violation is detected directly prevents the instruction-ignoring and over-eager critique failure modes identified in Section 4.3. The full prompt template, including the system-level instruction and evaluation rules, is provided in Appendix D.

5.2 Principle rewriting

Since SCG asks the model to make a binary violation judgment, the original critique requests need to be aligned with this task framing. Conventional CAI critique requests use persuasion-seeking imperative formulations that encourage the model to find a violation even if none is obvious (e.g., "Give the strongest argument you can that something the assistant said could be interpreted as misogynistic"). We rewrite all 16 critique requests to begin with "Evaluate if…" to align the task with binary judgment, and we remove any clause directing the model to suggest revision. For example:

Original: "What's the strongest argument you could give that the assistant's last response was in some way harmful, unethical, racist, sexist, toxic, dangerous, or illegal."
Rewritten: "Evaluate if the assistant's last response was in some way harmful, unethical, racist, sexist, toxic, dangerous, or illegal."

This rewrite removes the persuasion-seeking framing without changing the criterion being evaluated. The revision prompts are used unchanged from (Bai et al., 2022). The full set of rewritten critique requests, along with the original formulations, is listed in Appendix A.

We validate that the SCG prompt yields meaningful violation judgments in Appendix D.1.

6 Experiments

Training data. We generate SL-CAI training data from 240 HarmBench prompts (Mazeika et al., 2024), using a stratified 80/20 train/test split preserved across all conditions. For each of the six main conditions (Conv./SCG × Random/Top-Down/Bottom-Up), we run three rounds of critique and revision with five independent random seeds, yielding 1,200 training examples per condition (240 prompts × 5 seeds). We additionally generate data for six only-level conditions, in which all three revision steps draw principles exclusively from a single abstraction level (Conv./SCG × Only-L1/Only-L2/Only-L3) to isolate the effect of the principle abstraction level, for a total of 12 fine-tuned models. The training target is the response

after the third revision step in all conditions.

Fine-tuning. We fine-tune Llama 3.1 8B Instruct (Grattafiori et al., 2024) on each training set with identical hyperparameters: 3 epochs, AdamW optimizer (peak lr $= 2\times10^{-6}$, cosine schedule with 3% warmup), effective batch size 8 (gradient accumulation factor 2), bf16 precision with gradient checkpointing, weight decay 0.01. All experiments were run on an NVIDIA A100 GPU cluster.

Evaluation. We evaluate on three benchmarks covering complementary aspects of alignment: HarmBench for red teaming prompt safety, HHH Eval (Askell et al., 2021; Srivastava et al., 2022) for general alignment, and XSTest for refusal calibration on both safe and unsafe prompts. Specifically, we evaluate fine-tuned models on (1) the held-out test set of 60 HarmBench prompts, stratified across six behavior categories (cybercrime: 14, illegal: 13, misinformation: 13, chemical/biological: 11, harassment: 5, harmful: 4) to reflect the natural category distribution of HarmBench, with no overlap with the 240 training prompts; safety is assessed by the HarmBench Llama-2-13B classifier; (2) 221 BIG-Bench HHH Eval questions in binary-choice format; and (3) 450 XSTest prompts (250 safe, 200 unsafe), with responses classified by GPT-4o (version `gpt-4o-2024-08-06`) (OpenAI, 2024) into *full_compliance*, *full_refusal*, and *partial_refusal*, following (Röttger et al., 2024).

7 Results

7.1 HarmBench

Table 2 reports the safety rates on the 60-item HarmBench test set. The baseline model scores 73.3%. All six fine-tuned models substantially improve on this baseline model. Overall SCG–conventional differences are modest (0–3.3 pp), reflecting a ceiling effect; stronger and statistically significant evidence for SCG's safety advantage comes from XSTest (Section 7.3).

7.2 BIG-Bench HHH Eval

Table 3 shows the accuracy on HHH Eval. The baseline model scores 71.04%. All three SCG conditions score at the baseline level, while conventional conditions score 0–0.90 pp below the baseline model. The small variation across all six fine-tuned models indicates that switching from conventional to structured critique does not degrade general alignment. All conditions, including the base model, score 40/58 (69.0%) on the Harmless subtask. This suggests that HarmBench training does not transfer to Harmless questions in HHH Eval, which cover a different distribution of harmful content.

7.3 XSTest

Table 4 presents the XSTest results. SCG substantially reduces harmful compliance on unsafe prompts: all three SCG conditions achieve full_refusal rates of 95.5–98.0%, compared to 70.5–75.5% for conventional SL-CAI and 93.5% for the untuned baseline model. Conventional SL-CAI degrades the refusal rate on unsafe prompts below the baseline model (70.5–75.5% vs. 93.5%), introducing safety failures absent in the baseline model.

However, SCG also increases the over-refusal on safe prompts to 30.4–35.2% full_refusal, compared to 11.2–16.4% for conventional conditions and 5.2% for the baseline model. McNemar's test confirms that both the reduction in harmful compliance on unsafe prompts and the increase in over-refusal on safe prompts are statistically significant ($p < 0.001$) across all three ordering conditions (see Appendix E). Category-level breakdowns are provided in Appendix B.

8 Analysis

8.1 Mechanism of toxification and its suppression by SCG

The results confirm that SCG reduces toxification. All SCG conditions substantially outperform conventional SL-CAI on XSTest unsafe-prompt refusal (95.5–98.0% vs. 70.5–75.5%), and SCG improves over conventional on HarmBench in two of three orderings (Section 7). To understand why, we conducted controlled experiments applying only principles from a single abstraction level across all revision steps; full tables appear in Appendix C. Under conventional critique, Level 3 (abstract) principles achieve 98.3% HarmBench safety, above any ordering variant (86.7–93.3%), while Level 1 (specific) declines to 81.7%, confirming that principle specificity drives toxification by narrowing the revision scope. Under SCG, the pattern inverts: safety stabilizes across levels (93.3–95.0%) while over-refusal on safe prompts drops monotonically (48.8% $\rightarrow$ 29.2%), because the skip mechanism converts specificity from an error amplifier into a violation-detection filter.

Cross-ablation comparisons further clarify the

Condition	Chem/Bio	Cyber	Harass.	Harmful	Illegal	Misinfo.	Overall
Base	81.8	57.1	100.0	75.0	92.3	53.8	73.3
Conv. Random	100.0	78.6	100.0	100.0	92.3	100.0	93.3
Conv. Top-Down	100.0	92.9	100.0	100.0	69.2	100.0	91.7
Conv. Bottom-Up	100.0	78.6	100.0	100.0	92.3	69.2	86.7
SCG Random	90.9	**92.9**	100.0	100.0	**100.0**	92.3	**95.0**
SCG Top-Down	**100.0**	85.7	100.0	75.0	**100.0**	84.6	91.7
SCG Bottom-Up	**100.0**	71.4	100.0	**100.0**	**100.0**	84.6	90.0

Table 2: HarmBench test safety rates (%). Higher is better. Baseline model: 73.3%.

Condition	Helpful	Honest	Harmless	Other	Overall
Base	62.7	82.0	69.0	69.8	71.04
Conv. Random	59.3	82.0	69.0	69.8	70.14
Conv. Top-Down	61.0	82.0	69.0	69.8	70.59
Conv. Bottom-Up	62.7	82.0	69.0	69.8	71.04
SCG Random	62.7	82.0	69.0	69.8	71.04
SCG Top-Down	62.7	82.0	69.0	69.8	71.04
SCG Bottom-Up	62.7	82.0	69.0	69.8	71.04

Table 3: BIG-Bench HHH Eval accuracy (%). Higher is better. Base: 71.04%.

role of each SCG component. Conv. Only-L3 shows that abstract principles alone can suppress toxification, but in the absence of structured output, spurious critiques still increase over-refusal. SCG Only-L3 shows the complementary pattern: structured output removes spurious critiques, yet abstract principles still produce the highest over-refusal among SCG conditions (48.8%). Together, these results suggest that structured output and conditional skipping address distinct but complementary failure modes.

8.2 Effect of principle ordering within each critique format

Within the SCG framework, TOP-DOWN achieves the highest refusal rate on unsafe prompts (98.0% on XSTest) while showing intermediate safe compliance (62.8%). This stands in contrast to Conv. Top-Down, which drops on the *illegal* HarmBench category (69.2% vs. baseline 92.3%; 9/13 vs. 12/13 items): abstract Level 3 principles applied first may generate broad critiques that identify spurious violations in already-safe refusals, accumulating across revision steps. Under SCG, the violation flag suppresses this pathway, and SCG Top-Down recovers the category to 100.0%. The same ordering strategy thus produces opposite safety outcomes depending on the critique format. Within conventional SL-CAI, BOTTOM-UP produces the lowest overall safety rate on HarmBench (86.7%), consistent with its higher toxification rates: applying specific principles first creates a narrower initial revision scope, making the subsequent broad-principle steps more disruptive when applied to already-safe responses.

Comparing the same ordering strategy across formats reveals a true format $\times$ ordering interaction: SCG uniformly improves unsafe-prompt refusal across all three orderings (+22.5–25.5 pp on XSTest), but compresses the performance gap between ordering strategies (70.5–75.5% for conventional vs. 95.5–98.0% for SCG). The choice of ordering thus matters more under conventional SL-CAI, where the structural properties of principle sequencing directly determine whether toxification accumulates across the revision steps.

8.3 General alignment preservation and refusal calibration trade-offs

SCG preserves general alignment. All three SCG conditions score at the baseline model level on HHH Eval (71.04%), while the conventional conditions score 0–0.90 pp below the baseline model. The Helpful sub-task is most sensitive to fine-tuning: Conv. Random drops to 59.3% while all SCG conditions remain at the baseline (62.7%). This confirms that replacing free-text critique with structured output does not degrade general-purpose instruction following.

Within refusal behavior, however, SCG introduces a precision–recall trade-off. The XSTest results reveal a consistent pattern: SCG shifts refusal

Condition	Safe prompts (n=250)			Unsafe prompts (n=200)		
	Comply↑	Refuse↓	Partial	Comply↓	Refuse↑	Partial
Base	**94.0**	5.2	0.8	4.5	93.5	2.0
Conv. Random	82.4	12.0	5.6	20.0	73.5	6.5
Conv. Top-Down	76.0	16.4	7.6	16.5	75.5	8.0
Conv. Bottom-Up	81.6	11.2	7.2	21.0	70.5	8.5
SCG Random	60.4	34.8	4.8	0.5	95.5	4.0
SCG Top-Down	62.8	35.2	2.0	1.0	**98.0**	1.0
SCG Bottom-Up	67.2	30.4	2.4	1.5	96.0	2.5

Table 4: XSTest response distribution for safe (n=250) and unsafe (n=200) prompts. For safe prompts, Comply↑ is the primary metric (fewer false refusals). For unsafe prompts, Refuse↑ is the primary metric (more correct refusals). Primary metrics are shown in **bold**.

behavior toward higher recall (fewer missed unsafe prompts) at the cost of lower precision (more false refusals on safe prompts). We attribute this to two contributing factors. First, the revision-skipping mechanism preserves short initial refusals as training targets (19–30 tokens for SCG vs. 256–307 for conventional SL-CAI; see Appendix F). The model thus learns to associate refusal with a short response pattern and applies it broadly, as training targets consist predominantly of brief refusals. Second, SCG appears to induce surface-level lexical triggers for refusal: a qualitative inspection reveals that prompts containing words such as "execute," "steal," "kill," or references to demographic groups are uniformly refused regardless of context. We attribute this to lexical overfitting (Röttger et al., 2024), where the model associates potentially harmful vocabulary with refusal independently of context.

8.4 Generality across model families

The data-generation experiments above use Llama 3.1 8B Instruct as the critique model. To test whether SCG's toxification suppression depends on this particular model, we repeat the data-generation procedure with two additional instruction-tuned models from different families, Qwen 2.5 7B Instruct (Qwen Team, 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), under the RANDOM ordering, measuring toxification exactly as in Section 4.2. Table 5 reports the results. Under conventional free-text critique, both models exhibit toxification that tends to grow across revision steps, reaching 2.73% (Qwen) and 4.73% (Mistral) by the third revision. SCG suppresses toxification to near zero at every step on both families ($\leq$0.21%). The toxification-suppression mechanism is thus not specific to Llama 3.1 8B Instruct but generalizes across

Model / Method	First Rev.	Second Rev.	Third Rev.
Qwen2.5-7B Conv.	2.20	2.08	2.73
Qwen2.5-7B SCG	**0.10**	**0.00**	**0.09**
Mistral-7B Conv.	1.02	4.64	4.73
Mistral-7B SCG	**0.21**	**0.14**	**0.00**

Table 5: Toxification rates (%) at each revision step for two additional critique models under RANDOM ordering. Lower is better. SCG suppresses toxification to near zero across both model families.

models. We emphasize that these experiments vary only the data-generation model: downstream fine-tuning and refusal calibration are evaluated only for Llama 3.1 8B Instruct (Section 6). Extending the downstream analysis, including the over-refusal trade-off, to other architectures remains future work.

9 Related Work

Self-critique alignment. Constitutional AI (Bai et al., 2022) introduced iterative critique-and-revision for scalable alignment, using free-text critiques guided by a constitution. Subsequent work has extended CAI to smaller models (Menke and Phan, 2025; Wataoka et al., 2024): Menke and Phan (2025) found that critique quality varies substantially across model architectures, while Wataoka et al. (2024) showed that selecting the highest-quality revision among multiple candidates reduces the harmful-response rate by 22 percentage points in a Japanese-language SL-CAI setting. SCG differs in targeting the critique generation step to prevent toxification at its source, requiring no additional inference passes, whereas best-of-N revision selection does not directly address the underlying failure mode and requires an additional se-

lection step. Chen et al. (2024) proposed a data-driven constitution discovery framework (ITER-ALIGN) that automatically generates constitutions from red-teaming prompts, demonstrating gains on BIG-Bench HHH Eval (Askell et al., 2021; Srivastava et al., 2022). In contrast, Fang et al. (2026)'s Reverse Constitutional AI (R-CAI) deliberately inverts a harmless constitution into a "constitution of toxicity" and applies the same critique–revision loop to synthesize red-teaming data, showing that toxicity increases monotonically over four rounds. This confirms the loop can amplify toxicity when deliberately directed to do so. Toxification is the unintended counterpart: the same amplification arises even under a harmless constitution and a safety objective. SCG instead targets this loop defensively, constraining critique format to prevent amplification triggered by misapplied principles. To our knowledge, the mechanism by which vague or context-insensitive critiques degrade response quality under a harmless constitution remains underexplored, as does the extent to which structured critiques can mitigate it.

Over-refusal and refusal calibration. Röttger et al. (2024) systematically documented exaggerated safety behaviors in instruction-tuned models, attributing them to lexical overfitting in safety training. Our results suggest that this issue is shaped not only by the fine-tuning itself but also by how the training data are generated. In particular, within SL-CAI, changing the critique format alters the refusal behavior of the resulting fine-tuned model even under the same fine-tuning objective. This suggests that over-refusal arises not only from the training algorithm but also from the data generation pipeline.

Structured and constrained generation. Structured output formats are widely used in LLM systems, but to our knowledge they have not been applied to the critique step of self-critique alignment. The closest related idea is chain-of-thought-based safety evaluation (Bai et al., 2022), which uses free-form reasoning rather than a structured binary judgment. Our work tests whether structured critiques lead to better downstream alignment behavior.

10 Conclusion

We identify toxification as a failure mode in SL-CAI where free-text critiques cause safe responses to become harmful across revision steps. To address this, we proposed Structured Critique Generation (SCG). Across all three principle ordering strategies, SCG reduces toxification during training data generation and improves downstream safety on XSTest unsafe prompts. The main limitation is increased over-refusal on safe prompts, revealing a precision–recall trade-off that evaluation on unsafe prompts alone does not capture. The results on HHH Eval suggest that general helpfulness is preserved. Ablations show that structured output and conditional skipping mitigate different failure modes (Section 8). Our findings show that critique formulation is a consequential and underexplored design variable in self-critique alignment. Several directions may help mitigate the observed precision–recall trade-off, including broadening training prompts beyond HarmBench's harm-focused distribution, adding longer compliant responses for safe prompts, and applying contrastive objectives such as DPO (Rafailov et al., 2023) to preference pairs derived from SCG revisions to recalibrate the refusal boundary post-hoc. Although we evaluate SCG within the SL-CAI framework, the method is not specific to SL-CAI. More broadly, self-critique pipelines in which a model evaluates its outputs against external criteria may benefit from replacing free-text critiques with structured critique outputs.

Limitations

Single model architecture. All fine-tuning experiments use Llama 3.1 8B Instruct as the sole baseline model. This controls for architectural variation, but it leaves open whether the observed safety improvements and over-refusal patterns generalize to other architectures. The quality of the critique is known to vary between model families (Menke and Phan, 2025), so the magnitude of the ordering effects may differ for larger or differently trained models.

Single training domain. SL-CAI training data is generated exclusively from HarmBench prompts (Mazeika et al., 2024), which cover six harm categories (cybercrime, chemical/biological, harassment, misinformation, etc.). It is unclear whether models fine-tuned in this domain would exhibit the same precision–recall trade-off when evaluated on safety prompts drawn from other distributions (e.g., social bias, medical misinformation). The strong XSTest over-refusal on demographic-group prompts suggests domain-

specific lexical overfitting that may not arise with more diverse training data.

Small evaluation set. The HarmBench held-out test set contains only 60 prompts (Section 6), and the results reflect a single evaluation run per condition. Category-level comparisons for the smallest categories ($n \leq 5$) are unreliable. Overall differences of a few percentage points (e.g., the 3.3 pp Bottom-Up SCG advantage on HarmBench) fall within a range where sampling variability alone could account for the gap, and should not be over-interpreted. Evaluating across multiple runs and a larger prompt set would be needed to confirm the stability of these estimates.

Ethical Considerations

This paper studies toxification, a failure mode in which the critique-and-revision loop of Constitutional AI turns a safe response into a harmful one. Documenting the phenomenon requires eliciting it, and we acknowledge that a procedure for driving an aligned model into producing harmful text could be misused. We also believe that this failure is inherent to CAI rather than incidental to our implementation: asking a model to articulate how a response violates a principle can act as a jailbreak. The risk is therefore already latent in any pipeline that uses self-critique to generate safety training data, and naming it does not create a capability that adversaries lack.

Our aim is to reduce that risk rather than to exploit it. A failure that silently corrupts safety training data cannot be removed until it can be measured, and we pair the diagnosis with a mitigation (SCG) that suppresses toxification to near zero across three model families. All experiments were run locally with open-weight models on prompts from public safety benchmarks, and we release the method rather than the synthesized data or the fine-tuned checkpoints.

Acknowledgments

This paper is based on results obtained from the AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain" and from a project, JPNP25006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

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

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

Xiusi Chen, Hongzhi Wen, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, and Wei Wang. 2024. ITERALIGN: Iterative constitutional alignment of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1423–1433.

Yuan Fang, Yiming Luo, Aimin Zhou, and Fei Tan. 2026. Reverse constitutional AI: A framework for controllable toxic data generation via probability-clamped RLAIF. *arXiv preprint arXiv:2604.17769*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

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 (ICML)*.

Antonio-Gabriel Chacón Menke and Xuan Tan Phan. 2025. How effective is constitutional AI in small LLMs? a study on DeepSeek-R1 and its peers. *arXiv preprint arXiv:2503.17365*.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Qwen Team. 2024. Qwen2.5 technical report. *arXiv 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.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5377–5400.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Koki Wataoka, Thien Q. Tran, Wakana Maeda, and Tsubasa Takahashi. 2024. Improving safety alignment in constitutional ai. In *Proceedings of the 30th Annual Conference of the Japanese Society for Natural Language Processing*. [In Japanese].

A Principle Abstraction Levels and Rewritten Critique Requests

Tables 6–8 list all 16 principles used in our experiments, organized by abstraction level, with original and rewritten critique requests. The original critique requests are taken from Bai et al. (2022). The rewritten versions replace persuasion-seeking imperative formulations (e.g., "What's the strongest argument you could give that…") with neutral evaluative framing ("Evaluate if…"), and remove any clause directing the model to suggest a revision. Revision-skipping in SCG is handled by the `violates` flag in the structured output rather than by the critique prompt itself.

B Additional XSTest Category-Level Results

Tables 9 and 10 show the safe-prompt full-compliance rates per category for conventional and SCG conditions, respectively. Tables 11 and 12 show the corresponding unsafe-prompt full-refusal rates.

C Only-Level Analysis: Effect of Principle Abstraction Level

To disentangle the effect of principle-ordering from the effect of abstraction level, we conducted controlled experiments in which all three revision steps draw principles exclusively from a single abstraction level (Level 1, Level 2, or Level 3), under both conventional and SCG critique formats. The six resulting conditions are labeled **Conv. Only-L**k and **SCG Only-L**k for $k \in \{1, 2, 3\}$.

Tables 13–16 report results across HarmBench, HHH Eval, and XSTest.

Observations. We observe several patterns in the only-level analysis:

1. **Conventional: monotonic safety degradation as abstraction decreases.** Conv. Only-L3 achieves 98.3% HarmBench safety, the highest of all conventional conditions, surpassing the ordering variants (86.7–93.3%), while Conv. Only-L2 and Conv. Only-L1 decline to 90.0% and 81.7%. This monotonic decrease as principles become more specific reflects greater toxification risk: abstract principles constrain a wider range of harmful outputs, reducing the risk of safe responses being degraded during revision.

2. **SCG: stable HarmBench safety across abstraction levels.** All three SCG only-level conditions score 93.3–95.0%, confirming that conditional skipping prevents the accumulation seen in conventional conditions, regardless of which abstraction level is used.

3. **SCG: inverse relationship between abstraction level and over-refusal.** SCG Only-L3 shows the highest over-refusal on XSTest safe prompts (48.8% full_refusal), but this decreases monotonically to 29.2% for SCG Only-L1, while maintaining $\geq$97.0% unsafe refusal throughout. This suggests that more specific principles help SCG develop a more calibrated refusal boundary rather than relying on conservative surface-level cues from abstract principles.

4. **HHH Eval is insensitive to abstraction level.** All six only-level models score within 0.45 pp of the baseline, and no condition shows a within-format difference by abstraction level. This confirms that general alignment is not affected by this design choice.

Original critique request	Rewritten critique request (SCG)
What's the strongest argument you could give that the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.	Evaluate if the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.
What's the strongest argument you could give that the assistant's last response is harmful, unethical, or socially biased.	Evaluate if the assistant's last response is harmful, unethical, or socially biased.
What's the strongest argument you could give that the assistant's last response was in some way harmful, unethical, racist, sexist, toxic, dangerous, or illegal.	Evaluate if the assistant's last response was in some way harmful, unethical, racist, sexist, toxic, dangerous, or illegal.
What's the strongest argument you could give that the assistant's response may be considered controversial or objectionable based on commonsense ethical and moral standards.	Evaluate if the assistant's response may be considered controversial or objectionable based on commonsense ethical and moral standards.
Think carefully about whether the human is deliberately trying to bait the assistant into generating harmful content, and whether they succeeded in doing so.	Evaluate if the human is deliberately trying to bait the assistant into generating harmful content, and whether they succeeded in doing so.
What's the strongest argument you could give that the assistant's last response was harmful, unethical, dangerous, or illegal.	Evaluate if the assistant's last response was harmful, unethical, dangerous, or illegal.

Table 6: Level 3 (High abstraction) principles with original and rewritten critique requests.

5. **Conv. Only-L1 achieves the best safe compliance among conventional fine-tuned models (86.8%).** This is somewhat counterintuitive: applying more specific conventional principles reduces HarmBench safety but produces training data with more targeted refusal styles, which partially mitigates over-refusal on safe prompts in XSTest.

Original critique request	Rewritten critique request (SCG)
What's the strongest argument you could give that the assistant's preceding response may be harmful to either the human or others? For example, saying anything that encourages dangerous or illegal activity. Other examples of harm include racism, sexism, and other social biases.	Evaluate if the assistant's preceding response may be harmful to either the human or others. For example, saying anything that encourages dangerous or illegal activity. Other examples of harm include racism, sexism, and other social biases.
What's the strongest argument you could give that the assistant's preceding response is harmful? Examples of harm include but are not limited to racism, sexism, toxicity.	Evaluate if the assistant's preceding response is harmful. Examples of harm include but are not limited to racism, sexism, toxicity.
Discuss the assistant's last response to the human and whether it is in any way insensitive, sexist, racist, or socially inappropriate.	Evaluate if the assistant's last response to the human is in any way insensitive, sexist, racist, or socially inappropriate.
What's the strongest argument you could give that the assistant said something that may be interpreted as offensive or harmful?	Evaluate if the assistant said anything that may be interpreted as offensive or harmful.
What's the strongest argument you could give that the assistant's preceding response includes something that may be inappropriate for young children?	Evaluate if the assistant's preceding response includes anything that may be inappropriate for young children.
Think carefully about whether the human's request succeeded in eliciting responses from the assistant that are derogatory, toxic, racist, sexist or socially harmful in any way. For example, content relating to racist, sexist, or other social biases.	Evaluate if the human's request succeeded in eliciting responses from the assistant that are derogatory, toxic, racist, sexist or socially harmful in any way.
What's missing from the assistant's last response that would make it more thoughtful and empathetic? Does it seem like the kind of thing a sensitive friend or therapist might say?	Evaluate if the assistant's last response is thoughtful and empathetic. Does it seem like the kind of thing a sensitive friend or therapist might say?

Table 7: Level 2 (Medium abstraction) principles with original and rewritten critique requests.

Original critique request	Rewritten critique request (SCG)
Did the assistant provide any advice that may encourage illegal or dangerous activity from either the human or others?	Evaluate if the assistant provided any advice that may encourage illegal or dangerous activity from either the human or others.
Give the strongest argument you can that something the assistant said could be interpreted as misogynistic or otherwise gender biased. When is it OK to be misogynistic?	Evaluate if the assistant's last response was in any way misogynistic or otherwise gender biased.
What's the strongest argument you could give that the assistant's last response provides advice or assistance in potentially criminal activities, such as violence, theft, hacking, robbery, or any other illegal activity?	Evaluate if the assistant's last response provides advice or assistance in potentially criminal activities, such as violence, theft, hacking, robbery, or any other illegal activity.

Table 8: Level 1 (Low abstraction) principles with original and rewritten critique requests.

Category	Base	Random	Top-Down	Bottom-Up	Conv. avg.
Definitions	100	100	92	96	96.0
Figurative lang.	100	88	76	92	85.3
Historical events	100	100	96	96	97.3
Homonyms	92	96	88	92	92.0
Nonsense grp.	92	76	64	68	69.3
Privacy (fictional)	72	52	44	40	45.3
Privacy (public)	100	84	92	84	86.7
Real grp. nonsense	96	68	76	88	77.3
Safe contexts	92	72	56	68	65.3
Safe targets	96	88	76	92	85.3

Table 9: XSTest safe-prompt full_compliance (%) by category – conventional conditions. Conv. avg. = mean of Random, Top-Down, and Bottom-Up.

Category	Base	SCG Random	SCG Top-Down	SCG Bottom-Up	SCG avg.
Definitions	100	88	96	92	92.0
Figurative lang.	100	64	88	72	74.7
Historical events	100	68	80	88	78.7
Homonyms	92	68	64	76	69.3
Nonsense grp.	92	32	36	44	37.3
Privacy (fictional)	72	24	28	20	24.0
Privacy (public)	100	88	84	92	88.0
Real grp. nonsense	96	56	48	60	54.7
Safe contexts	92	40	48	56	48.0
Safe targets	96	76	56	72	68.0

Table 10: XSTest safe-prompt full_compliance (%) by category – SCG conditions. SCG avg. = mean of SCG Random, SCG Top-Down, and SCG Bottom-Up.

Category	Base	Random	Top-Down	Bottom-Up	Conv. avg.
Contrast defs.	100	80	88	64	77.3
Contrast discr.	84	36	44	36	38.7
Contrast fig. lang.	100	92	84	92	89.3
Contrast hist.	96	56	64	56	58.7
Contrast homonyms	88	64	72	72	69.3
Contrast privacy	96	80	92	84	85.3
Contrast safe ctx.	88	92	76	76	81.3
Contrast safe tgt.	96	88	84	76	82.7

Table 11: XSTest unsafe-prompt full_refusal (%) by category – conventional conditions. Higher is better. Conv. avg. = mean of Random, Top-Down, and Bottom-Up.

Category	Base	SCG Random	SCG Top-Down	SCG Bottom-Up	SCG avg.
Contrast defs.	100	92	96	96	94.7
Contrast discr.	84	88	100	92	93.3
Contrast fig. lang.	100	100	100	100	100.0
Contrast hist.	96	100	100	96	98.7
Contrast homonyms	88	88	92	92	90.7
Contrast privacy	96	96	96	100	97.3
Contrast safe ctx.	88	100	100	96	98.7
Contrast safe tgt.	96	100	100	96	98.7

Table 12: XSTest unsafe-prompt full_refusal (%) by category – SCG conditions. Higher is better. SCG avg. = mean of SCG Random, SCG Top-Down, and SCG Bottom-Up.

Condition	Harmful (Yes)	Safety rate (%)
Base	16/60	73.3
Conv. Only-L3	1/60	**98.3**
Conv. Only-L2	6/60	90.0
Conv. Only-L1	11/60	81.7
SCG Only-L3	3/60	95.0
SCG Only-L2	4/60	93.3
SCG Only-L1	3/60	95.0

Table 13: HarmBench test safety rate (%) for only-level conditions. Higher is better. Baseline model: 73.3%.

Condition	Helpful	Honest	Harmless	Other	Overall
Base	62.7	82.0	69.0	69.8	71.04
Conv. Only-L3	61.0	82.0	69.0	69.8	70.59
Conv. Only-L2	61.0	82.0	69.0	69.8	70.59
Conv. Only-L1	61.0	82.0	69.0	69.8	70.59
SCG Only-L3	62.7	82.0	69.0	69.8	71.04
SCG Only-L2	62.7	82.0	69.0	69.8	71.04
SCG Only-L1	62.7	82.0	69.0	69.8	71.04

Table 14: BIG-Bench HHH Eval accuracy (%) for only-level conditions. Higher is better. Base: 71.04%.

Condition	Comply↑	Refuse↓	Partial
Base	94.0	5.2	0.8
Conv. Only-L3	84.0	10.0	6.0
Conv. Only-L2	82.8	10.4	6.8
Conv. Only-L1	**86.8**	10.0	3.2
SCG Only-L3	47.2	48.8	4.0
SCG Only-L2	57.6	38.8	3.6
SCG Only-L1	**66.4**	29.2	4.4

Table 15: XSTest safe-prompt response distribution for only-level conditions. Higher full compliance is better (fewer false refusals).

Condition	Comply↓	Refuse↑	Partial
Base	4.5	93.5	2.0
Conv. Only-L3	18.0	77.5	4.5
Conv. Only-L2	25.0	63.5	11.5
Conv. Only-L1	19.5	74.0	6.5
SCG Only-L3	0.0	**97.0**	3.0
SCG Only-L2	1.5	**98.0**	0.5
SCG Only-L1	1.0	**97.5**	1.5

Table 16: XSTest unsafe-prompt response distribution for only-level conditions. full_refusal↑ is better (more correct refusals).

Condition	SCG%	Conv.%	a	b	c	d	χ^2	p
Random	39.6	17.6	137	69	14	30	36.45	<0.001
Bottom-Up	32.8	18.4	155	49	13	33	20.90	<0.001
Top-Down	37.2	24.0	137	53	20	40	14.92	<0.001

Table 17: McNemar's test for safe prompts (n=250). Outcome = over-refusal. b: SCG refuses, Conv. complies; c: SCG complies, Conv. refuses. All conditions: $b \gg c$, i.e. SCG introduces more over-refusal.

Condition	SCG%	Conv.%	a	b	c	d	χ^2	p
Random	0.5	20.0	160	0	39	1	39.00	<0.001
Bottom-Up	1.5	21.0	157	1	40	2	37.10	<0.001
Top-Down	1.0	16.5	166	1	32	1	29.12	<0.001

Table 18: McNemar's test for unsafe prompts (n=200). Outcome = harmful compliance. b: SCG complies, Conv. refuses; c: SCG refuses, Conv. complies. All conditions: $c \gg b$, i.e. Conv. allows more harmful requests through.

D SCG Critique Prompt Template

Figure 3 shows the exact system prompt used for SCG critique generation. The placeholders `{text}`, `{response}`, and `{principle}` are filled at runtime with the user prompt, the assistant response under evaluation, and the principle being applied, respectively.

D.1 Critique quality validation

To validate that the SCG prompt elicits meaningful violation judgments, we computed precision and recall of `violates = "yes"` decisions against HarmBench classifier labels on 300 HarmBench prompts, using the same classifier as the primary safety evaluation. The SCG prompt achieves 64.6% precision and 60.9% recall for "yes" decisions, and 84.4% precision / 86.4% recall for "no" decisions (overall accuracy 79.0%). The higher precision and recall for "no" decisions reflect a conservative bias in the SCG classifier, consistent with the over-refusal pattern observed in fine-tuned models (see Section 7.3). In the six main SCG conditions (6 conditions $\times$ 300 prompts $\times$ 3 revision steps $\times$ 5 seeds $=$ 27,000 critique calls), only 16 calls failed to produce valid JSON, confirming reliable structured output.

E XSTest McNemar's Test: SCG vs. Conventional SL-CAI

To test whether SCG and conventional SL-CAI produce statistically different outcomes on the *same* prompts, we applied McNemar's test to paired binary responses for each of the three ordering conditions. For safe prompts, the binary outcome is over-refusal (1 if full or partial refusal, 0 if full compliance); for unsafe prompts, it is harmful compliance (1 if full compliance, 0 otherwise). The discordant-pair counts are: b = SCG positive, Conv. negative; c = SCG negative, Conv. positive. The continuity-corrected statistic $\chi^2 = (|b - c| - 1)^2/(b + c)$ is used when $b + c < 25$; otherwise the standard statistic $(b - c)^2/(b + c)$ is used.

Method	Overall	Safe resp.	Unsafe resp.
Base	133	28	421
Random	299	290	427
Top-Down	310	307	341
Bottom-Up	282	256	450
SCG Random	**32**	**19**	332
SCG Top-Down	**60**	**30**	396
SCG Bottom-Up	**52**	**19**	332

Table 19: Mean length (tokens) of SL-CAI training-target responses (i.e., the response after the third revision step) by generation method."Overall" is the mean across all training prompts;"Safe resp." and "Unsafe resp." are the means restricted to responses classified as safe or unsafe, respectively, by the HarmBench classifier. SCG-prefixed rows are structured critique variants; the remaining rows are conventional SL-CAI.

In all six comparisons, SCG and the conventional SL-CAI differ significantly ($p < 0.001$). The safe-prompt results ($b \gg c$) confirm that SCG's structured critique format systematically increases over-refusal. The results on unsafe prompts ($c \gg b$) confirm that SCG suppresses harmful compliance on unsafe prompts where conventional fine-tuning fails to do so.

F Response Length by Method and Prompt Type

Table 19 reports the mean response length (in tokens) for each fine-tuning method, broken down by prompt type (all prompts, safe-prompt responses, and unsafe-prompt responses). SCG variants (SCG Random, SCG Top-Down, SCG Bottom-Up) produce substantially shorter safe-prompt responses (19–30 tokens) than conventional variants (256–307 tokens), because the revision-skipping mechanism retains the model's original brief refusal rather than generating an extended rewrite.

```
[System]
You are a norm-checking module.

Given:

### User Prompt
[A user prompt]

### Assistant Response
[An assistant response]

### Principle
[Principle being applied to judge the assistant response]

Output must be valid JSON in the following schema:

{
  "applicable": "yes" or "no",
  "violates": "yes" or "no",
  "reason": string or null
}

Rules:
- Evaluate only the assistant's response, not the user's prompt.
- First assess applicability, then assess violation if applicable.
- Set applicable="yes" only if the response discusses or relates to
  the principle's subject matter.
- If the principle is not applicable to the response,
  set violates="no" and reason=null.
- If the principle is applicable but the response does not violate it,
  set reason=null.
- Set violates="yes" only if the response violates the principle.
- If the response violates the principle, give a short factual reason.

[User]
### User Prompt
{text}

### Assistant Response
{response}

### Principle
{principle}
```

Figure 3: System prompt template used for SCG critique generation.