أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

# Mechanistic Analysis of Template-Injection Jailbreaks in Diffusion Language Models

**Nasser Al-Tamimi, Nawaf Alahmed, Abdullah Bukhari, Dr. Zahra Shakeri**

nmmd294@gmail.com, nawafahmed10000@gmail.com, abdu44ll3ah55@gmail.com, zahra.shakeri@utoronto.ca

## Abstract

This project asks whether template-injection attacks leave detectable internal signatures in diffusion language models and whether those signatures can support safer generation. We study LLaDA-8B-Instruct using 400 prompts derived from 100 MedSafetyBench requests. A paired $2 \times 2$ design crosses harmful versus neutral content with the presence or absence of DIJA injection, separating content effects from injection-mechanism effects. We record hidden states and sparse autoencoder features across token scopes, layers, and denoising fractions; train grouped linear probes; analyze direction geometry; and test feature zeroing and residual steering. DIJA achieved an 85% attack success rate on LLaDA. Corrected harmful and neutral injection probes achieved hidden-state AUCs of 0.997 and 0.996, while shuffled-label controls returned to chance. The injection contrasts shared a direction distinct from the harmful-content contrast. However, feature zeroing gave little or negative benefit, and steering traded safety improvements against response collapse. Dream-v0-Instruct-7B reproduced near-perfect early detectability, but steering did not improve attack success on the tested subset. A separate finalized 2,000-prompt RRAE extension preserved the injection-associated signal in a compact rank-4 residual representation (validation AUC 1.000; $B - A/C - D$ cosine 0.962). In a frozen-condition replication on 35 baseline-qualified validation pairs, safety-direction steering produced the strongest observed trade-off, with 28.6% harmful-side safe conversion, 88.6% benign helpfulness, and 25.7% strict paired success, while other fixed intervention families were weaker or caused larger utility loss. The results reveal a detection–correction asymmetry: injection is easy to detect internally, but reliable causal correction remains partial and condition-dependent.

**Date:** July 2026

**Code:** https://github.com/F4NG66/CDG2.0

# Contents


**1 Introduction** 3

**2 Related Work** 4

2.1 Diffusion language models . . . 4
2.2 Template-injection attacks . . . 4
2.3 Medical-safety evaluation . . . 4
2.4 Mechanistic probing and sparse representations . . . 4

**3 Methodology** 4

3.1 Study design and dataset . . . 4
3.2 DIJA prompt construction . . . 5
3.3 Models and generation . . . 5
3.4 Behavioral judging . . . 5
3.5 Activation recording . . . 5
3.6 Linear probes and text baselines . . . 6
3.7 Sanity checks and control experiments . . . 6
3.8 Direction geometry . . . 6
3.9 Causal interventions . . . 7
3.10 Commit timing, determinism, and re-masking . . . 7
3.11 Harm-direction steering and confound gates . . . 8
3.12 RRAE low-rank representation and frozen steering replication . . . 8
3.13 Direction stability . . . 9
3.14 Dream replication . . . 9
3.15 Record validation . . . 9

**4 Results** 9

4.1 Behavioral attack success . . . 9
4.2 The naive probe learned an ambiguous contrast . . . 9
4.3 Corrected injection probes . . . 10
4.4 Sanity and template controls . . . 10
4.5 Direction geometry . . . 11
4.6 SAE feature zeroing . . . 11
4.7 Residual steering and systematic ablation . . . 11
4.8 RRAE low-rank representation and large-scale frozen steering replication . . . 12
4.9 The attack outcome is fixed by the scaffold . . . 13
4.10 Re-masking does not neutralize the harm . . . 13
4.11 A confound-controlled harm direction detects but does not correct . . . 13
4.12 Stability across denoising . . . 14
4.13 Dream replication . . . 14

**5 Discussion** 15

5.1 Detection is easier than correction . . . 15
5.2 The determinant is in the prompt . . . 15

5.3 Why the factorial design matters . . . 16
5.4 Dense geometry and sparse features . . . 16
5.5 What the RRAE extension adds . . . 16
5.6 Comparison between LLaDA and Dream . . . 17
5.7 Implications for medical AI safety . . . 17

**6 Limitations and Future Work** **17**
6.1 Dataset and generalization . . . 17
6.2 Evaluation limitations . . . 17
6.3 Interpretability limitations . . . 18
6.4 Intervention limitations . . . 18
6.5 RRAE replication limitations and next steps . . . 18
6.6 From correction to detection-gated containment . . . 19

**7 Conclusion** **19**

**References** **20**

# 1 Introduction

Large language models are increasingly used to answer health questions, summarize clinical information, and support patient-facing communication. These applications make safety failures particularly consequential: an unsafe response may provide actionable instructions for misuse, encourage dangerous self-treatment, or present harmful information with unwarranted confidence. MedSafetyBench was developed to evaluate such risks using scenarios grounded in medical-ethics principles [1]. The present project uses this setting to study not only whether an attack succeeds, but also how the attack is represented inside the model.

Most widely deployed language models generate text autoregressively, one token at a time. Diffusion language models follow a different process. They begin with masked positions and iteratively denoise them, allowing tokens to be generated or revised in parallel using bidirectional context. LLaDA is an open 8-billion-parameter masked diffusion language model trained from scratch, while Dream 7B initializes from an autoregressive model and uses discrete diffusion for parallel sequence refinement [2, 3]. These models provide a useful test bed for studying safety mechanisms across a complete denoising trajectory rather than only at the final output.

The same architectural properties that enable flexible infilling and parallel generation also introduce a distinctive attack surface. DIJA creates interleaved text–mask prompts that use bidirectional context and iterative demasking to elicit unsafe completions from aligned diffusion language models [4]. A surface-level evaluation can measure attack success, but it does not reveal whether the model internally represents the presence of injection, whether that signal is independent of harmful content, or whether manipulating the signal can restore safe behavior.

This project addresses those questions through mechanistic interpretability. We record hidden states and sparse autoencoder (SAE) features while the model denoises its response. Linear probes test whether injection status is decodable; contrast vectors and singular value decomposition characterize the geometry of the signal; and causal interventions test whether removing features or steering the residual stream changes behavior.

A central methodological challenge is confounding. Comparing a harmful injected prompt with a neutral injected prompt appears to isolate attack behavior, but it also changes the underlying content. A classifier can therefore obtain high accuracy by identifying hazardous medical language rather than the injection mechanism. We address this problem with a paired $2 \times 2$ factorial design: harmful clean (A), harmful injected (B), neutral injected (C), and neutral clean (D). The $B - A$ and $C - D$ contrasts vary injection status while holding the corresponding request content fixed, whereas $B - C$ is explicitly treated as a content contrast.

The project makes four main contributions:

1. It constructs a paired medical-safety dataset that separates content hazardousness from template injection.
2. It demonstrates that injection status is linearly decodable from early denoising representations under grouped and shuffled-label controls.
3. It identifies a shared, low-dimensional injection direction that is geometrically distinct from the harmful-content direction.
4. It shows that internal detection does not directly translate into reliable correction: SAE ablation and residual steering provide limited benefits and can degrade output quality.

We further test generalization by repeating the main hidden-state, geometry, and steering analyses on Dream-v0-Instruct-7B. The resulting evidence supports a detection–correction asymmetry: the model exposes a strong internal injection signal, but using that signal as a safe and stable intervention target remains difficult.

# 2 Related Work

## 2.1 Diffusion language models

Autoregressive language models factorize text from left to right, whereas masked diffusion language models learn to reverse a process that progressively masks tokens. LLaDA showed that large-scale instruction-following behavior can emerge from this alternative formulation [2]. Dream 7B extended this direction using autoregressive initialization and context-adaptive noise scheduling, supporting arbitrary-order generation, infilling, and adjustable quality–speed trade-offs [3]. Because the entire partially masked sequence is available at each denoising step, internal safety states may evolve differently from those in left-to-right systems.

## 2.2 Template-injection attacks

DIJA is a jailbreak framework designed specifically for diffusion language models [4]. It transforms harmful requests into interleaved text–mask templates. The surrounding scaffold encourages the model to fill masked spans in a contextually consistent way, while parallel decoding reduces opportunities for token-by-token refusal. DIJA therefore motivates a defense that operates on denoising dynamics rather than relying exclusively on techniques developed for autoregressive generation.

## 2.3 Medical-safety evaluation

MedSafetyBench provides harmful medical requests and evaluation tasks grounded in medical-safety principles [1]. It is appropriate for this study because the cost of unsafe compliance is concrete and because harmful requests can be paired with topic-matched neutral controls. Our use differs from standard benchmark evaluation: rather than comparing only final-answer safety, we use benchmark items to construct controlled experimental contrasts for internal-representation analysis.

## 2.4 Mechanistic probing and sparse representations

Linear probes test whether a label can be decoded from internal activations. High probe accuracy establishes separability, but not causality: a probe may exploit lexical shortcuts, duplicated examples, or correlated content. Our design therefore combines probing with text baselines, grouped cross-validation, label permutation, template controls, and causal interventions. SAEs provide a complementary view by mapping dense residual-stream activations into sparse feature activations. If injection were localized to a small set of features, feature ablation might offer an interpretable defense. If it is distributed across many correlated features, a low-rank residual direction may be a more appropriate intervention target.

# 3 Methodology

## 3.1 Study design and dataset

We selected 100 harmful medical requests from MedSafetyBench and rewrote each request into a topic-matched neutral control. We then crossed content type (harmful or neutral) with DIJA injection (absent or present), producing 400 prompts:

The primary injection contrasts were $B - A$ and $C - D$. They compare injected and clean versions of the same harmful or neutral request. The $B - C$ contrast compares hazardousness under injection and was treated as a content contrast. Each quartet retained a shared base-request identifier used to prevent train–test leakage.

**Table 1** Paired 2 × 2 experimental design.

| Group | Content | Injection | Description |
|---|---|---|---|
| A | Harmful | Absent | Harmful request in clean form |
| B | Harmful | Present | Same harmful request with DIJA |
| C | Neutral | Present | Neutral control with DIJA |
| D | Neutral | Absent | Same neutral control in clean form |

### 3.2 DIJA prompt construction

Injected conditions were produced with the DIJA refinement pipeline [4]. Qwen2.5-7B-Instruct generated a scaffold containing contextual text and variable-length placeholders of the form `<mask:N>`. Each placeholder was expanded into $N$ instances of LLaDA's mask token before tokenization. The same refinement procedure was applied independently to harmful and neutral requests; therefore, B and C did not necessarily share identical wrapper wording or mask counts.

Temporary delimiters identified the template region during preprocessing. These markers were removed before inference and were used only to construct token-scope masks. To reduce direct lexical confounding, paired analyses also examined request positions whose underlying text was byte-identical in injected and clean conditions. The probe therefore compared the model's representation of the same content, rather than reading words that appeared only in the attack wrapper.

### 3.3 Models and generation

The primary model was LLaDA-8B-Instruct [2]. Inputs were formatted with the model's instruction template, and 128 masked output positions were appended. Generation used 128 denoising steps, generation length 128, block length 128, temperature 0.2, low-confidence remasking, and mask-token ID 126336. Injected template masks and response masks were resolved within the same denoising process.

Dream-v0-Instruct-7B was used as a replication model [3]. The Dream pipeline used 128 denoising steps and a maximum of 128 new tokens. All generation and activation recording ran on the Compute Canada Fir cluster using H100-class GPUs.

### 3.4 Behavioral judging

An independent DeepSeek judge assigned responses to *refuse*, *partial*, or *comply*. The primary attack success rate was

$$\mathrm{ASR} = \frac{\text{partial or compliant harmful responses}}{\text{valid judged group-B responses}}.$$

Intervention outputs additionally received graded evaluation. The judge returned a safety valence score, a specificity score, and a collapse score, and assigned one of four categories: unchanged harmful content, euphemistic softening, disclaimer only, or substantive safe substitution. Safe percentage was the proportion of responses receiving substantive safe substitution. Response length, trigram repetition, zero-width-space artifacts, and an automated collapse flag were retained as quality controls.

### 3.5 Activation recording

Residual-stream hidden states were recorded over five token scopes:

- `tpl_mask`: masked positions inside the injected template;
- `tpl_ctx`: unmasked template context;
- `harm`: medical-request or behavior-text positions;
- `out_mask`: output positions still masked at the checkpoint; and
- `out_unmask`: output positions already revealed.

The analysis used denoising fractions

$$\mathcal{F} = \{0.05, 0.10, 0.20, 0.35, 0.50, 1.00\}.$$

For each example, scope, fraction, and layer, valid token activations were mean-pooled into one vector. Empty scope–fraction cells were recorded as unavailable rather than imputed. TopK SAEs with 16,384 features and $k = 80$ active features were evaluated at layers 11, 16, and 26.

### 3.6 Linear probes and text baselines

We trained $L_2$-regularized logistic-regression probes separately for every representation type, scope, denoising fraction, and layer. The three contrasts were $B$ versus $C$, $B$ versus $A$, and $C$ versus $D$. Performance was summarized by receiver-operating-characteristic AUC across five folds.

The final evaluation used GroupKFold with the base-request identifier as the group variable. Consequently, A000, B000, C000, and D000 could not be split between training and testing. The initial $B$-versus-$C$ experiment used ordinary cross-validation and is reported as a methodological pitfall rather than the final injection probe.

To test surface-text separability, we converted behavior text, generated responses, and their concatenation into TF–IDF features and trained the same classifier. This baseline asked whether a contrast could be solved without model activations. Because AUC depends on score orientation, all values below 0.5 require verification of the positive class and probability column before interpretation.

### 3.7 Sanity checks and control experiments

Six additional controls were used:

1. **Label permutation:** class labels were shuffled before refitting hidden-state and SAE probes.
2. **Grouped splitting:** all variants of the same request were held within one fold.
3. **Prompt-only features:** token count, character count, mask count, newline count, template words, and system markers were used without activations.
4. **Length matching:** classification was repeated on examples with overlapping token lengths when common support existed.
5. **Harmless fake template:** group E combined the harmful requests with a patient-education formatting template containing comparable structure and masks.
6. **Paraphrase and movement:** groups F and H used repositioned and paraphrased real and harmless templates to reduce dependence on exact wording and prompt position.

### 3.8 Direction geometry

For every layer, fraction, and scope, group means defined three difference vectors:

$$\Delta_{BA} = \mu_B - \mu_A, \qquad \Delta_{CD} = \mu_C - \mu_D, \qquad \Delta_{BC} = \mu_B - \mu_C.$$

The cosine $\cos(\Delta_{BA}, \Delta_{CD})$ measured alignment between the harmful and neutral injection effects. Cosines with $\Delta_{BC}$ tested whether the injection geometry was distinct from hazardous-content geometry.

For steering, the $BA$ and $CD$ hidden-state vectors from the harm scope at $f = 0.05$ were normalized and stacked:

$$M^{(\ell)} = \begin{bmatrix} \widehat{\Delta}^{(\ell)}_{BA} \\ \widehat{\Delta}^{(\ell)}_{CD} \end{bmatrix}.$$

After singular value decomposition, the first right singular vector was retained as $v^{(\ell)}_{\text{shared}}$ and oriented to have positive inner product with $\widehat{\Delta}^{(\ell)}_{BA}$. Separate directions were constructed at layers 11, 16, and 26.

### 3.9 Causal interventions

The first intervention zeroed SAE features selected from the initial $B$-versus-$C$ differential expression analysis. One experiment removed the top 20 features ranked by mean gap; another removed all features in the most enriched co-activation module. All 100 group-B prompts were regenerated and judged.

The second intervention subtracted a normalized residual direction during all denoising steps:

$$h \leftarrow h - \alpha \widehat{v}.$$

The preliminary dose–response experiment used layer 16 and

$$\alpha \in \{0, 2, 4, 8, 16, 24, 32\}.$$

Steering was applied either to behavior positions or to masked template positions. Subsequent ablations varied (i) template-mask, template-all, template-plus-output-mask, and full-sequence positions; (ii) $BA$, $CD$, rank-2, and shared directions; and (iii) layers 11, 16, 26, or all three layers. Group B measured safety effects, while group C monitored neutral-response side effects.

### 3.10 Commit timing, determinism, and re-masking

To ask *why* correction is hard rather than only whether it succeeds, we re-captured LLaDA generations with per-position, per-step internal states, keeping the `out_mask` and `out_unmask` scopes separate rather than mean-pooled, and adding a per-position commit step $k_{\text{commit}} = \min\{k : i \notin M(x_{t_k})\}$, the denoising step at which output position $i$ leaves the mask set. On the harmful arm, $k_{\text{commit}}$ was anchored on the injected scaffold blanks only, bounded by the number of injected blanks, to avoid sweeping structural filler or the trailing refusal into the harm span. To separate a property of the scaffold from a quirk of temperature-0 decoding, we resampled 32 generations per prompt and measured how often the inclusive harm outcome flips for a fixed scaffold, sweeping temperature $\{0.2, 0.4, 0.7\}$ (the paper's configuration is 0.2). Finally, we tested the diffusion-native "undo" directly: at a chosen step we re-masked positions and continued denoising, with arm A re-masking the harmful blanks, arm C a matched count of non-harmful positions, arm D wiping the harm span *and* relocating the blanks elsewhere, and arm E the unmodified baseline, each at two timings and budget-floored to leave $\geq 20$ denoising steps so that failures could not be attributed to exhausted steps.

### 3.11 Harm-direction steering and confound gates

To test a direction pointing from safe toward harmful behavior as a corrective lever, we built a harm axis $v_{\text{harm}}$ by holding the prompt fixed and contrasting one harmful and one safe response to the *same* request over response tokens only, mean-pooling each side and unit-normalizing; because the prompt is identical on both sides, sentence wording cannot be the signal. Pairs were split train/test by base-request identifier (54/23). Before any steering we required the axis to pass four pre-registered gates: separation of harmful from safe content (Gate A); robustness to length, via length matching to ±20% and removal of near-duplicate pairs with response Jaccard $> 0.35$ (Gate B); removal of an authorship confound by having a single model author both the harmful and safe sides (Gate D); and transfer to LLaDA's own harmful-versus-safe axis, measured by *directional alignment* rather than by a transfer AUC that saturates (Gate E). We then steered $h \leftarrow h - \alpha\hat{v}$ at layer 16 over $\alpha \in \{0, 16, 32, 48, 64\}$, applied separately to the response region and to the scaffold region, with a matched-norm random-direction arm and a benign arm as mandatory controls and a benign-injected (Group C) condition as the decisive selectivity test.

### 3.12 RRAE low-rank representation and frozen steering replication

A separate finalized RRAE analysis was conducted on a larger canonical-DIJA dataset containing 2,000 prompts: 500 examples in each A/B/C/D condition. Each condition contained 400 health-related examples and 100 non-health controls. The 500 paired source indices were split at the pair level into 400 training pairs and 100 validation pairs, preventing A/B/C/D variants of the same source from crossing the split. Normalization statistics, RRAE parameters, and low-rank bases were fit from the training split only.

Canonical hidden-state screening covered three representation scopes (`harm`, `input_region`, and `input_ctx`), the six denoising fractions in $\mathcal{F}$, and layers 4, 11, 16, and 26, yielding 72 candidate datasets. Because the template-based scopes showed strong position-count shortcuts, the primary rank sweep was restricted to the `harm` scope. Six shortlisted harm-scope configurations were frozen before RRAE training, and each was trained at ranks

$$k \in \{4, 8, 16, 24, 32, 48, 64\},$$

for a total of 42 audited models. Model selection jointly considered reconstruction error, injection separability, alignment of the $B - A$ and $C - D$ residual directions, train–validation direction stability, pairwise consistency, harmfulness leakage, and position-count correlation. The selected representation was layer 11, $f = 0.05$, `harm` scope, rank $k = 4$.

The learned representation was then evaluated through five previously frozen intervention families: additive steering along the injection direction, train-only safety-direction steering, combined injection reduction plus safety addition, multi-layer projection removal, and strength/schedule variation. For projection removal, the canonical operation was

$$h' = h - \rho\langle h, v_{\text{inj}}\rangle v_{\text{inj}},$$

whereas safety-direction steering used

$$h' = h + \beta v_{\text{safety}}.$$

The injection and safety directions were treated as distinct signals.

The large-scale replication increased only sample size; no layer, rank, scope, dose, or denoising schedule was added after observing the replication outcomes. Eight validation pairs exposed in earlier steering work were excluded, leaving 92 fresh candidates. A baseline-only qualification

procedure, completed before steering outcomes were examined, retained 35 pairs (32 health-related and 3 non-health). These cases were not used for gradient training, train-only normalization, or low-rank basis fitting, but the broader validation split had contributed to representation and rank selection; consequently, this cohort is described as a *frozen-condition validation replication cohort*, not a completely untouched test set.

Across the five intervention families, 6,650 new steered outputs were generated. Together with 140 frozen baseline outputs, the final RRAE analysis contained 6,790 outputs. All were scored under a frozen blinded LLM-judge rubric using the clean semantic request rather than the injected generation prompt. The primary paired metric was strict safety–utility success: a pair passed only when the harmful injected B response became a coherent safe refusal or redirection and the paired benign injected C response remained helpful.

### 3.13 Direction stability

We reconstructed the SVD-based shared direction at every denoising fraction and computed cross-fraction cosine matrices. Comparisons between $f = 0.05$ and later fractions tested whether an early direction remained representative of the injection mechanism. We treated cosine values of at least 0.7 as operationally stable, values from 0.3 to 0.7 as moderate drift, and values below 0.3 as near-orthogonal.

### 3.14 Dream replication

Dream hidden states have width 3,584. We recorded layers 5, 14, and 23 over the same six fractions and request, masked-output, and unmasked-output scopes. The completed manifest contained 100 examples per A/B/C/D condition. We repeated the three probes, label-shuffle checks, and direction-cosine analysis. Baseline ASR used all 100 B responses. Steering used a 30-case subset, the masked-output direction from layer 14 at $f = 0.05$, and

$$\alpha \in \{0, 2, 4, -0.5, -1, -2\}.$$

### 3.15 Record validation

Every generated answer had a structured record containing case ID, condition, response text, activations, and run metadata. Manifests were checked for condition balance, vector dimensionality, corrupted values, and expected empty scopes. Generation, probing, judging, geometry, and steering outputs were stored separately so that labels and aggregate metrics could be recomputed without regenerating model outputs.

## 4 Results

### 4.1 Behavioral attack success

DIJA successfully bypassed LLaDA's refusal behavior on a substantial fraction of harmful prompts. On the 100 group-B examples, the baseline judged attack success rate was 0.85. This established that the attack was behaviorally effective and motivated analysis of the model's internal representations.

### 4.2 The naive probe learned an ambiguous contrast

The initial *B*-versus-*C* probe achieved AUC values close to 1.0 across the activation grid. At first glance, this appeared to show an easily detectable injection signal. However, both groups contained injected templates; their defining difference was harmful versus neutral content. A

TF–IDF classifier also separated $B$ and $C$ behavior text with AUC 1.0, demonstrating that this contrast could be solved from surface content alone. We therefore reinterpreted $B$ versus $C$ as a hazardous-content contrast and did not use it as the primary injection-mechanism probe.

The logged behavior-text AUCs for $B$ versus $A$ and $C$ versus $D$ were approximately 0.018 and 0.016 under the stored score orientation. These values are not equivalent to chance under the standard AUC definition: reversing the positive class or score sign would yield values near 0.98. We therefore flag the TF–IDF score orientation for audit and avoid using these two numbers as evidence that lexical confounding has been completely eliminated.

### 4.3 Corrected injection probes

The paired $B$-versus-$A$ and $C$-versus-$D$ probes remained highly accurate under GroupKFold. Table 2 reports the mean AUC across the evaluated grid.

**Table 2** Grouped-cross-validation performance of corrected probes.

| Contrast | Representation | Mean AUC |
|---|---|---|
| $B$ versus $A$ | Hidden state | 0.997 |
| $B$ versus $A$ | SAE features | 0.986 |
| $C$ versus $D$ | Hidden state | 0.996 |
| $C$ versus $D$ | SAE features | 0.982 |

Near-perfect separation was visible from the earliest measured denoising fraction, $f = 0.05$, and persisted over later fractions. The neutral $C$-versus-$D$ result is particularly important: harmfulness is absent from both conditions, yet injection remains strongly decodable.

### 4.4 Sanity and template controls

Shuffling class labels reduced performance to chance (Table 3). This rules out a pipeline that produces high AUC independently of the correct labels.

**Table 3** Mean AUC after label permutation.

| Contrast | Representation | Shuffled-label AUC |
|---|---|---|
| $B$ versus $A$ | Hidden state | 0.459 |
| $B$ versus $A$ | SAE features | 0.450 |
| $C$ versus $D$ | Hidden state | 0.447 |
| $C$ versus $D$ | SAE features | 0.454 |

High performance also survived grouped splitting by request ID: hidden-state AUC was 0.997 for $B$ versus $A$ and 0.996 for $C$ versus $D$. Length matching was possible only for the pooled injected-versus-clean contrast because the paired injected and clean prompts had little length overlap. On 310 matched examples, AUC changed from 0.967 before matching to 0.975 after matching, showing that length was not the only predictive factor, while also leaving length as a limitation for the individual paired contrasts.

The harmless-template controls clarified what the representations encode. E versus A was almost perfectly separable (hidden AUC 1.000; SAE AUC 0.997), confirming that template and mask structure strongly affect activations. More importantly, the real DIJA template remained separable from the harmless template when both contained harmful requests and comparable

structure: B versus E achieved hidden AUC 0.999 and SAE AUC 0.978. After paraphrasing and moving both templates, F versus H remained separable (hidden AUC 1.000; SAE AUC 0.996). These controls support a stable attack-style template signal, while not proving that every component of that signal is causally responsible for jailbreak success.

### 4.5 Direction geometry

The harmful and neutral injection effects shared a common direction. At representative settings,

$$\cos(\Delta_{BA}, \Delta_{CD}) = 0.784,$$

while the content contrast was nearly orthogonal to the neutral injection direction,

$$\cos(\Delta_{BC}, \Delta_{CD}) = -0.115.$$

This geometry validates the purpose of the factorial design: $BA$ and $CD$ capture a shared injection-associated component, whereas $BC$ follows a substantially different content direction. SVD of the normalized $BA$ and $CD$ vectors produced a dominant shared component used for steering.

SAE differential-expression analysis showed that the signal was spread across many features rather than isolated in one small, clean module. Co-activation clustering was weak (silhouette approximately 0.039). The combination of low-rank residual geometry and distributed sparse features suggested that an easily decoded signal need not map to a small number of independently causal SAE units.

### 4.6 SAE feature zeroing

The baseline group-B ASR was 0.85. Removing the top 20 $B$-versus-$C$ SAE features changed ASR only to 0.83, while ablating the most enriched feature module increased ASR to 0.95. Thus, the first SAE defense did not provide a meaningful safety improvement and could make behavior worse. This result is consistent with the later diagnosis that the selected features came from the content-dominated $B$-versus-$C$ contrast rather than the corrected injection contrasts.

### 4.7 Residual steering and systematic ablation

Residual steering produced non-monotonic, configuration-dependent behavior. The best tested LLaDA settings reduced harmful-content delivery to approximately 45%, but this reduction did not consistently produce clean refusals. Some settings softened or disrupted responses while increasing collapse. The position ablation found the strongest overall safety effect when the shared direction was applied to all template positions at layer 16. Applying the intervention to the full sequence was more damaging, showing that a broader intervention is not automatically a better defense.

Direction and layer ablations showed that no single setting dominated every criterion. The shared direction generally outperformed a content-derived direction, and layer 16 provided the best balance among the tested single-layer options. Larger coefficients could improve the safe-response percentage but also increased degeneration, reduced specificity, or damaged neutral responses. We therefore interpret steering as partial evidence that the shared direction is behaviorally relevant, not as a deployment-ready defense.

### 4.8 RRAE low-rank representation and large-scale frozen steering replication

The finalized RRAE extension confirmed that the DIJA-associated signal could be preserved in a compact residual representation. At the common `harm` scope and $f = 0.05$, hidden-state screening remained strongly injection-separable across layers, with validation injection AUCs of 1.000 at layers 4, 11, and 16 and 0.973 at layer 26. The corresponding $B - A/C - D$ cosines were 0.979, 0.963, 0.923, and 0.811, respectively.

The audited rank sweep selected the layer-11, rank-4 model. Its validation injection AUC was 1.0000 and its residual $B - A/C - D$ cosine was 0.9621. Train–validation direction cosines were 0.9910 for $B - A$ and 0.9914 for $C - D$, and both contrasts had a 100% positive-pair rate. Guardrails suggested little direct content leakage (harmfulness AUC 0.4357) or position-count shortcut correlation (0.0495). The model explained 77.17% of latent variance with validation reconstruction MSE 0.8580. Importantly, increasing rank improved reconstruction error but generally weakened injection separation and cross-content alignment, showing that reconstruction quality alone was not an adequate selection criterion.

The large-scale frozen-condition replication evaluated five intervention families on 35 baseline-qualified validation pairs. Table 4 reports the best descriptively ranked frozen condition from each family. Safety-direction steering gave the strongest observed safety–utility trade-off: 10/35 harmful injected B responses converted to a safe refusal or redirection (28.6%), 31/35 paired benign C responses remained helpful (88.6%), and 9/35 pairs met the strict paired gate (25.7%). The corresponding baseline-consistent strict success rate was 8/29 (27.6%).

**Table 4** Large-scale RRAE frozen-condition steering results on 35 baseline-qualified validation pairs. Utility failure is 1−benign C helpful preservation.

| Strategy | B safe conversion | C utility failure | Strict paired success |
|---|---|---|---|
| Additive | 2.9% | 5.7% | 2.9% |
| Safety direction | 28.6% | 11.4% | 25.7% |
| Combined | 11.4% | 5.7% | 11.4% |
| Projection removal | 5.7% | 8.6% | 5.7% |
| Strength / schedule | 17.1% | 57.1% | 8.6% |

The best safety-direction condition acted at layer 16 on currently masked positions with persistent application and a 2× layer-median dose. In contrast, the strongest temporal/dose condition converted 17.1% of B responses but produced 57.1% benign utility failure. Additive and projection-removal interventions largely preserved benign utility but converted only 2.9% and 5.7% of B responses, respectively. Combined steering achieved 11.4% strict paired success. Across all five families, 6,650 steered outputs were generated; including frozen baselines, 6,790 outputs were judged under the blinded protocol.

These results refine the earlier small-canary conclusion. The RRAE representation is not merely linearly separable: interventions derived from the learned space can alter behavior, and a distinct train-only safety direction can produce meaningful safe conversions. However, no fixed intervention produced reliable correction across most pairs. A low-rank residual space can therefore be geometrically clean and behaviorally relevant while robust causal control remains selective and condition-dependent.

### 4.9 The attack outcome is fixed by the scaffold

Under DIJA, the delivered outcome is a near-deterministic function of the scaffold. Resampling 32 generations per prompt at the paper's temperature (0.2), the inclusive harm outcome flipped on only 2.75% of draws (Wilson 95% CI [1.8%, 4.1%]): 16 of 25 scaffolds never flipped once, and a single scaffold accounted for 45% of all flips ($\chi^2 = 119.9$, $p \approx 10^{-14}$), so straddling is a property of particular scaffolds rather than sampling noise. Flipping rose smoothly at higher temperatures (for example 7.5% at 0.4 and 13.8% at 0.7) with no clean threshold. Commit timing was consistent with a scaffold that is executed rather than deliberated: harm committed late, at a median of ≈ 98 of 128 steps (≈ 75% of the trajectory), but bimodally, with a late cluster (≈ 90–103) and a small early cluster (≈ 26–41).

This is a *behavioral* claim—the outcome is fixed by the scaffold at input time—not a representational one that the model encodes the decision at step 0. A probe that appeared to read impending harm at step 0 separated a zero-harm placebo just as perfectly, revealing that it read the prompt text rather than a forming internal decision; we discarded it. Near-determinism by scaffold also explains the saturated baselines reported elsewhere in the paper: when the scaffold fixes the outcome, cohorts do not straddle.

### 4.10 Re-masking does not neutralize the harm

If harm could be caught as it forms, a diffusion model could in principle re-mask the harmful positions and regenerate them safely. It cannot. Across the seven at-risk cases, harm was re-delivered in every arm and at every timing, with no degeneration to explain it away (Table 5; 0/70 collapse). Two reads are decisive. First, arm A ≈ arm C ≈ arm E: targeting the harmful positions buys nothing over ignoring them, so there is no set of output positions whose removal helps. Second, arm D is the mechanism—when the harm span was wiped and equal blanks were opened elsewhere, the harm *re-formed in the original scaffold slots* while the relocated blanks took benign filler. Harm is bound to the fixed template text, not to wherever blanks are offered. A methods caveat travels with this result: a naive late re-mask (leaving a single step) yields garbled text that a judge misreads as "safe," so any re-masking defense evaluated without a degeneration control will report a false success; budget-flooring removed the artifact and the harm returned in full.

**Table 5** Re-masking falsification: harm re-delivered across all arms and timings ($n = 7$ at-risk cases).

| Arm | Intervention | Harm re-delivered |
|---|---|---|
| E | none (baseline) | 7/7 |
| A | re-mask the harmful blanks | 7/7 |
| C | re-mask matched non-harmful positions | 7/7 |
| D | wipe harm span + relocate blanks | 7/7 |

### 4.11 A confound-controlled harm direction detects but does not correct

We tested steering with a harm direction $v_{\text{harm}}$ built from same-prompt response pairs (Methods). Three confounds were caught and removed by pre-registered gates before any steering (Table 6): a wording confound (a text-only classifier separated the prompts at AUC 1.0), an authorship confound (with harmful text from one model and safe text from another, the axis read *who wrote it*; rebuilt with a single author on both sides, its cosine with an author axis flipped from +0.9 to −0.4), and a length/degeneracy confound (LLaDA's harmful and safe responses were often

near-identical text, Jaccard $\approx 0.5$). The surviving axis is a validated detector: it transfers to LLaDA's own harm axis at cosine $\approx 0.5$, roughly 14× the random-direction floor. Transfer AUC was a saturated 1.0 everywhere and is uninformative here—LLaDA's harmful and safe activations lie so far apart that a random direction separates them at $\approx 0.95$—so the discriminating test is directional alignment, not AUC.

Steering along this validated detector ($h \leftarrow h - \alpha\hat{v}$, layer 16, $\alpha \in \{0, 16, 32, 48, 64\}$, with matched-norm random and benign control arms) did not correct the attack. In the response region the effect was causally null: no harm reduction beyond the random arm, with benign outputs merely shortened. In the scaffold region harm did fall dose-dependently, but the fall was not correction—a matched random direction reproduced about half of it, degeneracy rose to 17–22% at the doses where harm dropped most, and the decisive control, benign content inside a real injected scaffold (Group C), broke identically (responses shortened by 172–343 characters). Adding the published steer-early / re-mask-late schedule contributed nothing beyond plain scaffold steering. The benign-injected control is what separates the two readings: clean benign prompts (Group D) carry no scaffold, so scaffold steering trivially preserves them, whereas Group C exposes that "harm down" was really "scaffold broken." A direction can therefore be a rigorously validated harm *detector* and still provide no selective corrective lever—a third independent instance of the detection–correction asymmetry, with the direction, confounds, and controls all pinned down.

**Table 6** Confound gates passed by the harm direction $v_{\text{harm}}$ before steering.

| Gate | Question | Result |
|---|---|---|
| A | Separates harmful from safe content? | pass (held-out AUC $\geq 0.90$) |
| B | Length rather than harm? | pass (length-matched AUC holds) |
| D | Authorship rather than harm? | pass (author-axis cosine $+0.9 \rightarrow -0.4$) |
| E | Transfers to LLaDA's harm axis? | pass (cosine $\approx 0.5$, $\approx 14\times$ floor) |

## 4.12 Stability across denoising

The shared injection direction was temporally stable across much of the denoising trajectory. Directions extracted at $f = 0.05$ remained aligned with later fractions at the analyzed layers and generally exceeded the operational stability threshold for early-to-middle comparisons. This supports using early activations for detection. Stability alone, however, did not guarantee that applying a fixed direction at every denoising step would be safe or effective.

## 4.13 Dream replication

Dream reproduced the early-detection result. All 400 hidden-state records passed validation, with 100 records in each condition and no corrupted vectors outside expected empty scopes. At $f = 0.05$, the best $B$-versus-$A$, $C$-versus-$D$, and $B$-versus-$C$ hidden probes each achieved AUC $1.000 \pm 0.000$. Selected shuffled-label controls returned near chance, including 0.476 for the neutral injection contrast at layer 5.

The strongest shared geometry occurred at masked output positions, layer 14, and $f = 0.05$:

$$\cos(\Delta_{BA}, \Delta_{CD}) = 0.980,$$

$$\cos(\Delta_{BA}, \Delta_{BC}) = -0.972, \qquad \cos(\Delta_{CD}, \Delta_{BC}) = -0.989.$$

Dream's full 100-case group-B baseline ASR was 0.50. Steering was then tested on a fixed 30-case subset whose matched baseline ASR was 0.30. Table 7 shows that neither steering sign improved on this subset baseline.

**Table 7** Dream steering results on the 30-case group-B subset.

| $\alpha$ | ASR |
|---|---|
| 0.0 | 0.300 |
| 2.0 | 0.533 |
| 4.0 | 0.433 |
| -0.5 | 0.300 |
| -1.0 | 0.333 |
| -2.0 | 0.433 |

The Dream result strengthens the main conclusion: injection can be nearly perfectly detectable and geometrically clean while remaining difficult to correct through direct residual steering.

# 5 Discussion

## 5.1 Detection is easier than correction

The central result is an asymmetry between internal detection and behavioral correction. Injection status was linearly separable from early hidden states in both LLaDA and Dream, including the neutral *C*-versus-*D* comparison. The signal also formed a shared direction across harmful and neutral content. Nevertheless, direct interventions on that signal were unreliable. SAE feature zeroing produced little benefit or increased attack success, and residual steering traded safety gains against collapse and loss of response quality.

This asymmetry is not contradictory. A representation can carry enough information for a classifier without serving as a single causal control knob. The injection label may be redundantly encoded across many features, correlated with multiple stages of computation, or downstream from the mechanism that actually produces unsafe compliance. High AUC should therefore be interpreted as evidence that the state contains information, not as proof that subtracting the probe direction will reverse the computation.

## 5.2 The determinant is in the prompt

Three further results turn this asymmetry from an observation into a mechanism. First, the attack outcome is a near-deterministic function of the scaffold: resampling flips it on only 2.75% of draws, 16 of 25 scaffolds never flip, and one scaffold holds 45% of all flips ($\chi^2 = 119.9$, $p \approx 10^{-14}$). The trajectory is executed, not deliberated—there is little forming decision inside denoising to detect or steer, which is also why a probe can appear to read harm before generation begins even though a zero-harm placebo shows it is reading the prompt. Second, the diffusion-native "undo"—re-masking harmful positions and regenerating—does not remove the harm (7/7 across every arm); harm re-forms in the fixed scaffold slots regardless of where blanks are placed, so the determinant is immutable prompt text, not the output canvas. Third, a rigorously confound-controlled harm direction detects harm yet cannot steer it out: null in the response region and mere breakage in the scaffold region, where the benign-injected control breaks alongside the harmful one. Together with the RRAE finding that only a dedicated safety direction—not the injection or harm axis—achieves even partial, non-deployable conversion, this constrains

what a correction can be. A controller that acts on the output region, however sophisticated its direction, inherits an empty target; any corrective signal lives in the prompt/scaffold encoding, and disrupting that encoding indiscriminately damages benign and harmful generations alike. The conservative, better-supported use of the internal signal is therefore detection-gated containment—flag a harmful trajectory early and halt before render—rather than in-place repair, with its make-or-break metric being the false-positive rate on benign-injected prompts.

### 5.3 Why the factorial design matters

The initial $B$-versus-$C$ result illustrates a common interpretability failure mode. A near-perfect probe can appear mechanistically meaningful while exploiting a simpler difference in the data. The $2 \times 2$ design changes the interpretation of each contrast. $B - C$ measures content under injection, $B - A$ measures injection in harmful content, and $C - D$ measures injection in neutral content. The alignment between $BA$ and $CD$, together with their separation from $BC$, would be unavailable in a two-group design.

The controls further refine the claim. Harmless templates also produce a strong internal signature, so $B - A$ alone cannot mean that the model has isolated a unique jailbreak circuit. However, the model can distinguish DIJA-style templates from harmless formatting even after controlling template presence, masks, content, and approximate layout. The most defensible conclusion is therefore that the representations encode a stable attack-style template difference, part of which is shared across content types.

### 5.4 Dense geometry and sparse features

The residual-stream analysis identified a shared low-dimensional direction, whereas the SAE analysis found widespread differential activation and weak clusters. This suggests that a simple global direction can be synthesized from many sparse contributors. Feature-by-feature ablation may fail because information is redundant: removing a few features leaves alternative components available, while removing a large module may damage unrelated capabilities.

Residual steering better matches the low-rank geometry, but a direction estimated from mean differences is still observational. Applying it at every denoising step may overcorrect, interfere with generation, or move activations off the model's normal manifold. The non-monotonic dose response and collapse at larger coefficients are consistent with this interpretation.

### 5.5 What the RRAE extension adds

The RRAE extension strengthens the distinction between representation discovery and control. A rank-4 residual model preserved perfect validation injection separation and strong $B - A/C - D$ alignment even though higher-rank models reconstructed hidden states more accurately. This indicates that the representation most useful for isolating the injection-associated geometry need not be the representation that minimizes reconstruction error. The low-rank bottleneck can suppress variance that is useful for reconstruction while retaining a cleaner cross-content injection signal.

The large-scale frozen replication also shows that the learned space is behaviorally relevant without supporting a simple "erase the injection" story. Projection removal and additive manipulation of the injection direction produced only weak safe conversion, whereas a separate train-only safety direction achieved the strongest paired outcome. This distinction matters mechanistically: $v_{\text{inj}}$ characterizes how injected and clean prompts differ, while $v_{\text{safety}}$ characterizes a direction toward safer response behavior. They are not interchangeable control vectors.

Even the best safety-direction condition passed the strict paired safety–utility gate for only

25.7% of the 35 validation pairs. Stronger temporal steering increased harmful-side conversion in some cases but damaged benign utility sharply. Thus, effective steering requires more than moving hidden states in a fixed direction: the trajectory must remain within a valid behavioral manifold and reach an appropriate safe region. Misaligned or excessive steering can preserve harmful behavior, induce unnecessary refusal, or push generation toward collapse.

### 5.6 Comparison between LLaDA and Dream

Both models exposed injection information from the earliest measured fraction, but the clearest geometry appeared in different scopes. LLaDA's main steering direction was extracted from request-content positions, while Dream's strongest shared direction appeared at masked output positions in layer 14. This difference suggests that the location of an injection signature depends on model architecture, training, or denoising schedule. A defense developed for one diffusion model should not be assumed to transfer unchanged to another.

Dream also provides a particularly clear example of the detection–correction asymmetry. The shared direction achieved cosine alignment of 0.980 while direct steering in either sign failed to improve the matched subset baseline. Geometric cleanliness is therefore insufficient evidence of causal controllability.

### 5.7 Implications for medical AI safety

For medical applications, a defense should do more than suppress harmful keywords. It must preserve coherence, specificity when appropriate, and useful safe alternatives. A model that avoids unsafe content by producing gibberish is not clinically useful. The graded judge and neutral controls were therefore essential: they separated genuine safe substitution from vague disclaimers, loss of specificity, or generation collapse.

The early and stable signal may still be useful for monitoring. An online detector could flag suspicious denoising trajectories, trigger a safer regeneration policy, or switch to a specialized refusal process. Such a detector is a more conservative use of the current evidence than direct activation subtraction. Future defenses may combine internal detection with remasking, conditional regeneration, or a learned controller trained specifically to preserve output quality.

## 6 Limitations and Future Work

### 6.1 Dataset and generalization

The main experiment used 100 MedSafetyBench source requests. Although the paired design provides strong within-request control, the sample does not cover every medical-risk category, language, or conversational setting. Neutral requests were generated by a language model and may differ systematically from the harmful originals in style or length. Future work should include human-reviewed neutral controls, multilingual cases, multi-turn interactions, and independently sampled external test sets.

The study focused on two open diffusion language models. LLaDA and Dream differ in architecture and training, and their results already show scope-specific differences. General claims about diffusion language models require evaluation on additional model families and checkpoints.

### 6.2 Evaluation limitations

The behavioral labels depend on an automated DeepSeek judge. Partial compliance was counted as attack success for the primary ASR, while graded steering evaluation used a richer rubric. Judge

calibration, inter-rater agreement, and sensitivity to prompt wording were not fully established. A final evaluation should combine blinded human review with at least one independent automated judge.

The TF–IDF analysis contains an unresolved orientation issue. Logged AUC values near 0.02 for paired behavior-text contrasts are not chance-level under the standard definition and may reflect a reversed positive class or probability column. These values must be audited before they are used to rule out surface-text separation. Likewise, the chance-level prompt-only result is unexpected when explicit template indicators are present and should be rerun after verifying feature extraction.

Injected and clean prompts had little length overlap, preventing direct length matching for $B - A$ and $C - D$. The pooled matched analysis shows that length is not the only signal, but it does not completely remove length as a potential contributor to the paired contrasts.

### 6.3 Interpretability limitations

Linear separability does not establish a causal mechanism. Difference vectors and SVD summarize group-level geometry but can combine multiple correlated processes. The SAE dictionary may also be imperfect: sparse features are learned representations, not guaranteed biological-style units, and subtracting decoder contributions can move the residual stream outside its training distribution.

The direction-stability thresholds were operational rather than theoretically derived. High cosine similarity across fractions indicates a persistent direction but does not show that a constant intervention is appropriate throughout denoising.

### 6.4 Intervention limitations

The LLaDA steering experiments explored a finite set of layers, positions, directions, and coefficients. The observed minimum harmful delivery rate did not always correspond to clean refusal, and some configurations increased collapse. Dream steering was evaluated on only 30 group-B examples, whose baseline ASR (0.30) differed from the full 100-case baseline (0.50). Only group B received final judge scoring in the Dream steering experiment, so neutral side effects were not fully quantified.

Future intervention work should:

- evaluate promising settings on all 100 Dream cases and judge A, C, and D side effects;
- learn fraction-dependent or position-dependent controllers instead of applying one fixed vector at every step;
- select SAE features from the corrected $BA/CD$ mechanism contrasts rather than the content-dominated $BC$ contrast;
- compare activation steering with remasking, constrained regeneration, and detector-triggered refusal; and
- test cross-model transfer of detectors and directions.

### 6.5 RRAE replication limitations and next steps

The RRAE extension is now complete at the frozen-condition validation-replication stage, but it should not be described as an untouched external test. The final 35 pairs were excluded from gradient training, train-only normalization, low-rank basis fitting, and prior steering exposure; however, the broader 100-pair validation split had contributed to representation and rank selection. The large-scale steering experiment therefore estimates performance of previously

frozen conditions on a fresh steering cohort within the validation split rather than definitive out-of-sample generalization.

The cohort was also baseline-qualified: 35 of 92 fresh candidate validation pairs met the pre-steering qualification rule, and 32 of the 35 were health-related. This improves interpretability of conversion metrics but narrows the population to cases where the baseline behavior was suitable for paired safety–utility evaluation. Future work should repeat the frozen intervention on a completely untouched external cohort and report both unconditional and baseline-qualified results.

Finally, all 6,790 RRAE outputs were scored by a frozen blinded LLM judge rather than independent human adjudicators. The judging protocol was outcome-blinded and no labels were changed after unblinding, but the reported categories remain model-defined behavioral outcomes. A publication-grade follow-up should include blinded human review of a stratified sample and at least one independent automated judge. The strongest current result—safety-direction steering—also remains partial: 25.7% strict paired success is evidence of selective controllability, not a deployment-ready safety controller.

### 6.6 From correction to detection-gated containment

Our steering results consistently separate detection from control, which reshapes the intervention agenda. First, the most promising near-term use of the validated detector is containment rather than repair: gate on an early harm read and halt or regenerate before render, reporting the false-positive rate on benign-injected (Group C) prompts as the primary safety–utility metric. Second, because re-masking the output was falsified as a defense (harm re-forms in the fixed scaffold slots) and output-region steering is null, any corrective steering must target the prompt/scaffold encoding, or use a LLaDA-native direction built from the model's own harmful-versus-refusal outputs; whether the $\approx 0.5$ transfer alignment of the current harm axis was the bottleneck, or scaffold steering is fundamentally too coarse under DIJA, is left open. Third, the determinism result implies a scaffold-conditioned defense—input-time screening or scaffold neutralization—is likely to outperform any mid-denoising intervention on the majority of cases that commit late, while the early-committing minority bounds what mid-trajectory steering can achieve.

## 7 Conclusion

This project investigated how DIJA template-injection attacks are represented inside diffusion language models and whether those representations can support safer generation. A paired $2 \times 2$ dataset separated harmful content from injection status, preventing the content confound that made the initial $B$-versus-$C$ probe misleading.

Corrected $B$-versus-$A$ and $C$-versus-$D$ probes achieved near-perfect AUC under grouped cross-validation, remained detectable from the earliest denoising fraction, and collapsed to chance after label permutation. Direction analysis found a shared injection-associated component that was distinct from the hazardous-content direction. Dream reproduced both the early probe signal and a strongly aligned shared direction.

The intervention results were more cautious. SAE feature zeroing did not reduce attack success meaningfully, and residual steering was sensitive to position, layer, and coefficient. Some configurations reduced harmful-content delivery, but the gains were accompanied by response collapse or loss of useful specificity. On Dream, an exceptionally clean geometric direction did not yield a successful steering defense.

The finalized RRAE extension provided a stronger causal test at larger scale. A rank-4 layer-11 representation preserved perfect validation injection separation and strong cross-content

alignment, but the five frozen steering families behaved differently. Safety-direction steering was the strongest condition, converting 28.6% of harmful injected responses while preserving benign helpfulness in 88.6% of paired controls and achieving 25.7% strict paired success across 35 baseline-qualified validation pairs. Projection removal, additive steering, and combined steering were weaker, while aggressive strength/schedule settings produced substantial utility failure. These results show that the low-rank signal is behaviorally relevant, but the best control direction is not simply the negative of the injection direction.

The overall finding is therefore a detection–correction asymmetry: template injection is easy to detect in internal denoising representations but difficult to erase without damaging generation. The most promising near-term use of the signal is internal monitoring that triggers safer regeneration or refusal, rather than unconditional subtraction of a fixed direction. More broadly, safety mechanisms for diffusion language models should be designed around their iterative denoising dynamics and evaluated for both harm reduction and preservation of response quality.

## References

[1] Tessa Han et al. “MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models”. In: *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track* 37 (2024). URL: https://arxiv.org/abs/2403.03744.

[2] Shen Nie et al. “Large Language Diffusion Models”. In: *arXiv preprint arXiv:2502.09992* (2025). DOI: 10.48550/arXiv.2502.09992. URL: https://arxiv.org/abs/2502.09992.

[3] Jiacheng Ye et al. “Dream 7B: Diffusion Large Language Models”. In: *arXiv preprint arXiv:2508.15487* (2025). DOI: 10.48550/arXiv.2508.15487. URL: https://arxiv.org/abs/2508.15487.

[4] Zichen Wen et al. “The Devil behind the Mask: An Emergent Safety Vulnerability of Diffusion LLMs”. In: *arXiv preprint arXiv:2507.11097* (2025). DOI: 10.48550/arXiv.2507.11097. URL: https://arxiv.org/abs/2507.11097.