Novelty Assessment Report

Paper: AlphaSteer: Learning Refusal Steering with Principled Null-Space Constraint
PDF URL: https://openreview.net/pdf?id=1vvbzAqdTe
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

As LLMs are increasingly deployed in real-world applications, ensuring their ability to refuse malicious prompts, especially jailbreak attacks, is essential for safe and reliable use. Recently, activation steering has emerged as an effective approach for enhancing LLM safety by adding a refusal direction vector to internal activations of LLMs during inference, which will further induce the refusal behaviors of LLMs. However, indiscriminately applying activation steering fundamentally suffers from the trade-off between safety and utility, since the same steering vector can also lead to over-refusal and degraded performance on benign prompts. Although prior efforts, such as vector calibration and conditional steering, have attempted to mitigate this trade-off, their lack of theoretical grounding limits their robustness and effectiveness. To better address the trade-off between safety and utility, we present a theoretically grounded and empirically effective activation steering method called AlphaSteer. Specifically, it considers activation steering as a learnable process with two principled learning objectives: utility preservation and safety enhancement. For utility preservation, it learns to construct a nearly zero vector for steering benign data, with the null-space constraints. For safety enhancement, it learns to construct a refusal direction vector for steering malicious data, with the help of linear regression. Experiments across multiple jailbreak attacks and utility benchmarks demonstrate the effectiveness of AlphaSteer, which significantly improves the safety of LLMs without compromising their general capabilities. Our codes are available at \url{https://anonymous.4open.science/r/AlphaSteer-929C/}.

Core Task Landscape

This paper addresses: **learning refusal steering for large language model safety enhancement**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Refusal Mechanism Analysis and Representation**
- **Refusal Steering and Control Methods**
- **Over-Refusal Mitigation and Utility Preservation**
- **Safety Evaluation and Benchmarking**
- **Alternative Safety Paradigms and Response Strategies**
- **Domain-Specific and Multimodal Safety**
- **Safety Infrastructure and Guardrails**
- **Adversarial Robustness and Attack Surfaces**
- **Reinforcement Learning and Preference Optimization for Safety**

Complete Taxonomy Tree

- learning refusal steering for large language model safety enhancement Survey Taxonomy
- Refusal Mechanism Analysis and Representation
  - Refusal Direction Identification and Characterization (5 papers)
  - [5] Refusal in language models is mediated by a single direction (Andy Arditi, 2024) View paper
  - [31] Llms encode harmfulness and refusal separately (Zhao Jia-chen, 2025) View paper
  - [40] COSMIC: Generalized Refusal Direction Identification in LLM Activations (Vincent Siu, 2025) View paper
  - [43] Refusal Behavior in Large Language Models: A Nonlinear Perspective (Maier, 2025) View paper
  - [45] Refusal Direction is Universal Across Safety-Aligned Languages (Wang Xinpeng, 2025) View paper
  - Safety Layer and Component Localization (3 papers)
  - [1] On prompt-driven safeguarding for large language models (Zheng, 2024) View paper
  - [13] Safety Layers in Aligned Large Language Models: The Key to LLM Security (Li Shen, 2024) View paper
  - [37] Understanding Refusal in Language Models with Sparse Autoencoders (Yeo Wei Jie, 2025) View paper
  - Harmfulness Encoding and Conceptual Separation (1 papers)
  - [25] Finding and Reactivating Post-Trained LLMs' Hidden Safety Mechanisms (M Li, 2025) View paper
- Refusal Steering and Control Methods
  - Activation-Based Steering Techniques ★ (9 papers)
  - [0] AlphaSteer: Learning Refusal Steering with Principled Null-Space Constraint (Anon et al., 2026) View paper
  - [6] Automating steering for safe multimodal large language models (Wang Mengru, 2025) View paper
  - [26] SafeSwitch: Steering Unsafe LLM Behavior via Internal Activation Signals (Peixuan Han, 2025) View paper
  - [30] Steering without side effects: Improving post-deployment control of language models (Stickland, 2024) View paper
  - [32] Internal activation as the polar star for steering unsafe llm behavior (Qian Cheng, 2025) View paper
  - [46] Feature-Guided SAE Steering for Refusal-Rate Control using Contrasting Prompts (Zhu, 2025) View paper
  - [48] SARSteer: Safeguarding Large Audio Language Models via Safe-Ablated Refusal Steering (Lin, 2025) View paper

    - [49] Scaling laws for activation steering with Llama 2 models and refusal mechanisms (Sheikh Abdur Raheem Ali, 2025) View paper
    - [50] LatentGuard: Controllable Latent Steering for Robust Refusal of Attacks and Reliable Response Generation (Shu Huizhen, 2025) View paper
    - Training-Based Refusal Enhancement (7 papers)
    - [3] Rule based rewards for language model safety (Mu Tong, 2024) View paper
    - [9] Safety is Not Only About Refusal: Reasoning-Enhanced Fine-tuning for Interpretable LLM Safety (Zhang Yu-you, 2025) View paper
    - [11] Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning (Chen Si, 2025) View paper
    - [21] Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training (Youliang Yuan, 2025) View paper
    - [22] Safety pretraining: Toward the next generation of safe ai (Maini, 2025) View paper
    - [39] HumorReject: Decoupling LLM Safety from Refusal Prefix via A Little Humor (Wu, 2025) View paper
    - [41] Robust LLM safeguarding via refusal feature adversarial training (Yu Lei, 2024) View paper
    - Controllable and Adaptive Safety Alignment (3 papers)
    - [4] Refusal tokens: A simple way to calibrate refusals in large language models (Jain, 2024) View paper
    - [23] Safeconstellations: Steering llm safety to reduce over-refusals through task-specific trajectory (Yadav Sumit, 2025) View paper
    - [29] Controllable safety alignment: Inference-time adaptation to diverse safety requirements (Zhang Jingyu, 2024) View paper
- Over-Refusal Mitigation and Utility Preservation
    - Over-Refusal Analysis and Characterization (3 papers)
    - [27] When Safety Blocks Sense: Measuring Semantic Confusion in LLM Refusals (Riad Ahmed Anonto, 2025) View paper
    - [35] Understanding and Mitigating Over-refusal for Large Language Models via Safety Representation (Junbo Zhang, 2025) View paper
    - [47] Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary (Pan, 2025) View paper
    - Over-Refusal Reduction Techniques (2 papers)
    - [15] Refusal-Aware Red Teaming: Exposing Inconsistency in Safety Evaluations (Yongkang Chen, 2025) View paper
    - [17] Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning (Zhang, 2025) View paper
- Safety Evaluation and Benchmarking
    - Refusal Behavior Evaluation Benchmarks (2 papers)
    - [2] Sorry-bench: Systematically evaluating large language model safety refusal (Xie, 2024) View paper
    - Comprehensive Safety Assessment Frameworks (4 papers)
    - [8] Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms (Han, 2024) View paper
    - [28] From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law (John Mavi, 2025) View paper
    - [38] Should LLM Safety Be More Than Refusing Harmful Instructions? (Dras, 2025) View paper
    - [42] SafeLawBench: Towards Safe Alignment of Large Language Models (Zhu Han, 2025) View paper
- Alternative Safety Paradigms and Response Strategies (2 papers)
    - [12] Backtracking improves generation safety (Zhang Yiming, 2024) View paper
    - [18] From hard refusals to safe-completions: Toward output-centric safety training (Yuan Yuan, 2025) View paper
- Domain-Specific and Multimodal Safety
    - Multimodal Safety and Unlearning (2 papers)
    - [7] Refusing Safe Prompts for Multi-modal Large Language Models (Liu Hongbin, 2024) View paper
    - [14] Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning (Chen Junkai, 2025) View paper
    - Agent and Contextual Safety (1 papers)
    - [19] Refusal-trained llms are easily jailbroken as browser agents (Kumar, 2024) View paper
- Safety Infrastructure and Guardrails (2 papers)
    - [16] Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment (Olga Shvetsova, 2025) View paper
    - [33] A Voter-Based Stochastic Rejection-Method Framework for Asymptotically Safe Language Model Outputs (Jake R. Watts, 2024) View paper
- Adversarial Robustness and Attack Surfaces (1 papers)
    - [34] The devil behind the mask: An emergent safety vulnerability of diffusion llms (Wen Zichen, 2025) View paper
- Reinforcement Learning and Preference Optimization for Safety (4 papers)
    - [10] A Minimalist Approach to LLM Reasoning: from Rejection Sampling to Reinforce (Xiong Wei, 2025) View paper
    - [20] Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback (Xu, 2024) View paper
    - [24] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models (Khaki, 2024) View paper
    - [36] Rule based rewards for fine-grained llm safety (T Mu, 2024) View paper

Narrative

Core task: learning refusal steering for large language model safety enhancement. The field has organized itself around several complementary branches that together address how models decide to refuse harmful requests and how those decisions can be improved. One major branch focuses on understanding refusal mechanisms through representation analysis, examining how models internally encode safety-related features and decision boundaries. Another branch develops steering and control methods that directly manipulate model activations or apply targeted interventions to guide refusal behavior, with works like Refusal Direction[5] and AlphaSteer[0] exploring activation-based techniques. A third branch tackles over-refusal mitigation, seeking to preserve utility when safety measures become too conservative, while parallel efforts concentrate on evaluation benchmarks such as Sorry Bench[2] and domain-specific safety challenges. Additional branches explore alternative response strategies beyond simple refusal, adversarial robustness against jailbreaks, and reinforcement learning approaches that optimize safety through preference data.

Within the activation-based steering cluster, a particularly active line of work investigates how to extract and apply refusal-related directions in model representation space. AlphaSteer[0] sits squarely in this area, emphasizing methods that steer model behavior by intervening on internal activations during inference. Nearby approaches like SafeSwitch[26] and Steering Without Side Effects[30] share

this focus on activation manipulation but differ in their treatment of trade-offs: some prioritize minimizing unintended impacts on model capabilities, while others explore how to make steering more robust or interpretable. A contrasting thread examines whether refusal can be localized to specific layers or features, with Feature Guided SAE[46] and SARSteer[48] probing the granularity of safety representations. The central tension across these methods involves balancing effective refusal of genuinely harmful requests against maintaining helpfulness on benign queries, a challenge that motivates ongoing exploration of how steering vectors generalize and whether they can be applied selectively without degrading overall performance.

Related Works in Same Category

The following **8 sibling papers** share the same taxonomy leaf node with the original paper:

1. Automating steering for safe multimodal large language models

Authors: Wang Mengru, Lyucheng Wu, Xu, Ziwen, Mengru Wang, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent progress in Multimodal Large Language Models (MLLMs) has unlocked powerful cross-modal reasoning abilities, but also raised new safety concerns, particularly when faced with adversarial multimodal inputs. To improve the safety of MLLMs during inference, we introduce a modular and adaptive inference-time intervention technology, AutoSteer, without requiring any fine-tuning of the underlying model. AutoSteer incorporates three core components: (1) a novel Safety Awareness Score (SAS) that a...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to induce refusal behaviors in language models. They overlap in using activation-level interventions during inference to enhance safety without model retraining, and both address the trade-off between safety and utility preservation. The key difference is that AlphaSteer learns a transformation matrix with null-space constraints for benign prompts and refusal direction reconstruction for malicious prompts, while AutoSteer uses a safety prober to detect toxicity and conditionally triggers a refusal head, specifically targeting multimodal large language models with automated layer selection via Safety Awareness Score.

---

2. SafeSwitch: Steering Unsafe LLM Behavior via Internal Activation Signals

Authors: Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Heng Ji, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) exhibit exceptional capabilities across various tasks but also pose risks by generating harmful content. Existing safety mechanisms, while improving model safety, often lead to overly cautious behavior and fail to fully leverage LLMs'internal cognitive processes. Inspired by humans'reflective thinking capability, we first show that LLMs can similarly perform internal assessments about safety in their internal states. Building on this insight, we propose SafeSwitch, a...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to control refusal behavior in LLMs. They overlap in using internal activation signals to induce refusal for malicious prompts while preserving utility on benign ones. However, AlphaSteer learns a transformation matrix constrained to the null space of benign activations to dynamically construct steering vectors, whereas SafeSwitch employs a safety prober to monitor internal states and conditionally activates a specialized refusal head only when unsafe content is predicted, representing different mechanisms for achieving selective steering.

---

3. Steering without side effects: Improving post-deployment control of language models

Authors: Stickland, Asa Cooper, Lyzhov, Alexander, Asa Cooper Stickland, et al. (13 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Language models (LMs) have been shown to behave unexpectedly post-deployment. For example, new jailbreaks continually arise, allowing model misuse, despite extensive red-teaming and adversarial training from developers. Given most model queries are unproblematic and frequent retraining results in unstable user experience, methods for mitigation of worst-case behavior should be targeted. One such method is classifying inputs as potentially problematic, then selectively applying steering vectors o...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to control refusal behavior in LLMs. They overlap in addressing the safety-utility trade-off when applying steering vectors to induce refusal, with both methods aiming to preserve model capabilities on benign inputs while enhancing safety on malicious ones. The key difference is that AlphaSteer learns a transformation matrix constrained to the null space of benign activations to dynamically construct steering vectors, while this candidate paper (Steering without side effects) uses KL-divergence minimization to train models to be robust against the side effects of pre-computed steering vectors, then applies conditional steering based on input classification.

---

4. Internal activation as the polar star for steering unsafe llm behavior

Authors: Qian Cheng, Peixuan Han, Chen, Xiusi, Cheng Qian, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) exhibit exceptional capabilities across various tasks but also pose risks by generating harmful content. Existing safety mechanisms, while improving model safety, often lead to overly cautious behavior and fail to fully leverage LLMs' internal cognitive processes. Inspired by humans' reflective thinking capability, we first show that LLMs can similarly perform internal assessments about safety in their internal states. Building on this insight, we propose SafeSwitch,...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to control refusal behavior in LLMs. They overlap in using internal activation signals to induce refusal for malicious prompts while preserving utility on benign ones. However, AlphaSteer learns a transformation matrix constrained to the null space of benign activations to construct steering vectors dynamically, whereas SafeSwitch employs a safety prober to monitor internal states and conditionally activates a specialized refusal head only when unsafe content is predicted, representing different mechanisms for achieving selective activation steering.

---

5. Feature-Guided SAE Steering for Refusal-Rate Control using Contrasting Prompts

Authors: Zhu, Zining, Samaksh Bhargav, Zining Zhu | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Model (LLM) deployment requires guiding the LLM to recognize and not answer unsafe prompts while complying with safe prompts. Previous methods for achieving this require adjusting model weights along with other expensive procedures. While recent

advances in Sparse Autoencoders (SAEs) have enabled interpretable feature extraction from LLMs, existing approaches lack systematic feature selection methods and principled evaluation of safety-utility tradeoffs. We explored using differen...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations during inference to control refusal behavior in LLMs. While AlphaSteer learns a transformation matrix with null-space constraints to dynamically construct steering vectors that preserve utility on benign prompts while inducing refusal on malicious ones, the candidate paper uses Sparse Autoencoders (SAEs) to identify and steer specific interpretable features through a contrasting prompt methodology. The key difference is that AlphaSteer employs principled null-space projection for utility preservation with learned refusal direction reconstruction, whereas the candidate paper focuses on systematic SAE feature selection and amplification/suppression strategies guided by composite scoring of differential activations.

---

6. SARSteer: Safeguarding Large Audio Language Models via Safe-Ablated Refusal Steering

Authors: Lin, Weilin, Li, Jianze, Weilin Lin, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Audio-Language Models (LALMs) are becoming essential as a powerful multimodal backbone for real-world applications. However, recent studies show that audio inputs can more easily elicit harmful responses than text, exposing new risks toward deployment. While safety alignment has made initial advances in LLMs and Large Vision-Language Models (LVLMs), we find that vanilla adaptation of these approaches to LALMs faces two key limitations: 1) LLM-based steering fails under audio input due to t...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to induce refusal behaviors in language models. While AlphaSteer addresses safety enhancement in text-based LLMs through null-space-constrained steering that learns to construct refusal vectors for malicious prompts while preserving benign activations, SARSteer targets Large Audio-Language Models (LALMs) by deriving refusal steering from text modality and applying safe-space ablation to handle audio inputs. The key distinction is that AlphaSteer operates on text-based LLMs with a principled null-space learning framework, whereas SARSteer extends activation steering to the multimodal audio domain with text-derived vectors to address modality-specific challenges.

---

7. Scaling laws for activation steering with Llama 2 models and refusal mechanisms

Authors: Sheikh Abdur Raheem Ali, Justin Xu, Ivory Yang, Jasmine Xinze Li, Ayse Arslan, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

As large language models (LLMs) evolve in complexity and capability, the efficacy of less widely deployed alignment techniques are uncertain. Building on previous work on activation steering and contrastive activation addition (CAA), this paper explores the effectiveness of CAA with model scale using the family of Llama 2 models (7B, 13B, and 70B). CAA works by finding desirable'directions'in the model's residual stream vector space using contrastive pairs (for example, hate to love) and adding ...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to control refusal behavior in LLMs. They overlap in their core approach of adding refusal direction vectors to model activations during inference to induce safety-enhancing refusal responses. However, the original paper (AlphaSteer) introduces a learnable transformation matrix with principled null-space constraints to preserve utility on benign prompts while steering malicious ones, whereas the candidate paper empirically investigates scaling laws of contrastive activation addition (CAA) across different Llama 2 model sizes (7B, 13B, 70B) without proposing a novel steering mechanism or addressing the utility-safety trade-off through theoretical constraints.

---

8. LatentGuard: Controllable Latent Steering for Robust Refusal of Attacks and Reliable Response Generation

Authors: Shu Huizhen, Li Xuying, Huizhen Shu, Li Zhuo, Xuying Li, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Achieving robust safety alignment in large language models (LLMs) while preserving their utility remains a fundamental challenge. Existing approaches often struggle to balance comprehensive safety with fine-grained controllability at the representation level. We introduce LATENTGUARD, a novel three-stage framework that combines behavioral alignment with supervised latent space control for interpretable and precise safety steering. Our approach begins by fine-tuning an LLM on rationalized dataset...

Relationship Analysis

Both papers belong to the Activation-Based Steering Techniques category, focusing on modifying internal activations to control refusal behavior in LLMs. They overlap in their core approach of manipulating hidden representations during inference to induce refusal for malicious prompts while preserving utility on benign ones. However, AlphaSteer learns a transformation matrix constrained to the null space of benign activations using linear regression, while LatentGuard employs a supervised variational autoencoder (VAE) with multi-label annotations to learn disentangled latent representations, offering a probabilistic framework with explicit semantic supervision rather than null-space projection.

Contributions Analysis

Overall novelty summary. The paper introduces AlphaSteer, a theoretically grounded activation steering method that applies learnable transformations to refusal direction vectors during inference to enhance LLM safety. It resides in the Activation-Based Steering Techniques leaf, which contains nine papers including the original work. This leaf sits within the broader Refusal Steering and Control Methods branch, indicating a moderately populated research direction focused on inference-time interventions. The taxonomy shows this is an active area with multiple concurrent approaches exploring how to manipulate internal activations to induce refusal behaviors without modifying model weights.

The taxonomy reveals that Activation-Based Steering Techniques is one of three sibling categories under Refusal Steering and Control Methods, alongside Training-Based Refusal Enhancement (seven papers) and Controllable and Adaptive Safety Alignment (three papers). Neighboring branches include Refusal Mechanism Analysis and Representation, which studies internal refusal structures, and Over-Refusal Mitigation and Utility Preservation, which addresses the safety-utility trade-off from a diagnostic perspective. The paper's focus on learnable steering with utility preservation connects it to over-refusal concerns while remaining distinct from training-based approaches that modify weights or adaptive frameworks that adjust safety thresholds dynamically.

Among 22 candidates examined across three contributions, no clearly refuting prior work was identified. The AlphaSteer method itself examined six candidates with zero refutable matches, the learnable transformation mechanism examined six candidates with zero refutations, and the null-space projection technique examined ten candidates with zero refutations. This suggests that within the limited search scope of top-K semantic matches and citation expansion, the specific combination of theoretical grounding, learnable

transformations, and null-space constraints for utility preservation appears relatively novel. However, the analysis explicitly notes this is not an exhaustive literature search, and the moderate density of the parent leaf indicates active parallel work in activation steering.

Based on the limited search scope of 22 candidates, the work appears to occupy a distinct position within a moderately crowded research direction. The absence of refuting candidates across all three contributions suggests novelty in the specific technical approach, though the taxonomy structure shows the broader activation steering paradigm is well-established with eight sibling papers. The analysis does not cover exhaustive prior work in adjacent areas like training-based methods or adaptive alignment, which may contain relevant comparisons not captured by semantic search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: AlphaSteer: a theoretically grounded activation steering method with null-space constraints

Description: The authors propose AlphaSteer, a novel activation steering approach that uses null-space constraints to preserve utility on benign prompts while learning to construct refusal direction vectors for malicious prompts. This method addresses the safety-utility trade-off through principled learning objectives rather than heuristic designs.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A Multi-Task Energy-Aware Impedance Controller for Enhanced Safety in Physical Human-Robot Interaction

URL: View paper

Brief Assessment

Energy Aware Impedance[72] addresses robot impedance control with null-space constraints for physical safety in human-robot interaction, not LLM activation steering or safety-utility trade-offs in language models.

---

2. Pixel: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration

URL: View paper

Brief Assessment

Pixel[67] focuses on position-wise adaptive steering with dual-view subspace construction and closed-form intervention strength optimization, rather than null-space constraints for safety-utility trade-offs. The technical approaches differ fundamentally in their optimization objectives and constraint formulations.

---

3. MEUV: Achieving Fine-Grained Capability Activation in Large Language Models via Mutually Exclusive Unlock Vectors

URL: View paper

Brief Assessment

MEUV[71] focuses on decomposing refusal directions into topic-specific, mutually exclusive vectors for fine-grained capability unlocking in security settings. AlphaSteer addresses a different problem: preserving utility on benign prompts while enhancing safety against jailbreaks through null-space constraints. The technical approaches and objectives differ fundamentally.

---

4. MOSAICO: offline synthesis of adaptation strategy repertoires with flexible trade-offs

URL: View paper

Brief Assessment

MOSAICO[70] focuses on offline synthesis of adaptation strategy repertoires with flexible trade-offs in system adaptation scenarios, not on activation steering methods for LLM safety with null-space constraints.

---

5. Msrs: Adaptive multi-subspace representation steering for attribute alignment in large language models

URL: View paper

Brief Assessment

Msrs[68] focuses on multi-attribute steering through subspace representation fine-tuning with orthogonal subspace allocation, while AlphaSteer addresses safety-utility trade-offs using null-space constraints for single-attribute refusal steering. The technical approaches and application domains differ substantially.

---

6. What makes and breaks safety fine-tuning? a mechanistic study

URL: View paper

Brief Assessment

Safety Fine Tuning Mechanics[69] focuses on analyzing how safety fine-tuning methods transform MLP weights to align unsafe inputs into null space, rather than proposing a new activation steering method with null-space constraints for the safety-utility trade-off.

Contribution 2: Learnable activation steering mechanism with transformation matrix

Description: The authors introduce a learnable transformation matrix that dynamically constructs steering vectors based on prompt activations, enabling data-driven and fine-grained control over the steering process instead of relying on fixed vectors or manual thresholds.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Steering LLMs' Reasoning With Activation State Machines

URL: View paper

Brief Assessment

Activation State Machines[56] focuses on dynamic state-space modeling for reasoning trajectories using recurrent state updates (predict-correct cycles), while the original paper learns a transformation matrix for constructing steering vectors with null-space constraints for safety-utility preservation. These are distinct technical approaches to activation steering.

---

2. Fusion Steering: Prompt-Specific Activation Control

URL: View paper

Brief Assessment

Fusion Steering[54] focuses on prompt-specific activation control for factual accuracy in QA tasks using dynamic injection weights optimized per prompt, rather than a learnable transformation matrix for general refusal steering as in the original paper.

---

3. Understanding Reasoning Mechanisms in Large Language Models Through Direction Learning

URL: View paper

Brief Assessment

Direction Learning[55] focuses on understanding reasoning mechanisms through direction learning in LLMs, not on learnable transformation matrices for activation steering in safety contexts. The candidate's sparse context does not provide evidence of a similar dynamic steering mechanism for refusal behavior.

---

4. FairSteer: Inference Time Debiasing for LLMs with Dynamic Activation Steering

URL: View paper

Brief Assessment

FairSteer[53] focuses on fairness/debiasing through fixed steering vectors computed from contrastive prompt pairs, not learnable transformation matrices that dynamically construct steering vectors based on prompt activations as in the original paper.

---

5. DPD-LoRA: Dynamic Prompt-Driven Low-Rank Adaptation for Improved Generalization

URL: View paper

Brief Assessment

DPD LoRA[52] focuses on low-rank adaptation for vision-language models using prompt-driven guidance, not on activation steering for LLM safety. The transformation matrices in DPD-LoRA serve a fundamentally different purpose—adapting model parameters for downstream tasks rather than dynamically constructing steering vectors based on prompt activations for refusal behavior.

---

6. DynaGuide: Steering Diffusion Polices with Active Dynamic Guidance

URL: View paper

Brief Assessment

DynaGuide[51] focuses on steering diffusion policies for robot control using dynamics models during the denoising process, not on activation steering with transformation matrices for LLM prompt-based control.

Contribution 3: Null-space projection for utility preservation

Description: The authors develop a null-space projection method that constrains the steering transformation to produce near-zero vectors for benign prompts, ensuring their activations remain unchanged and thus preserving model utility on non-harmful tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Privacy by Projection: Federated Population Density Estimation by Projecting on Random Features

URL: View paper

Brief Assessment

Privacy by Projection[66] addresses federated population density estimation using null-space projection to preserve privacy in location data, not LLM activation steering for safety. The technical domains and applications are entirely different.

---

2. Falcon: Fine-grained activation manipulation by contrastive orthogonal unalignment for large language model

URL: View paper

Brief Assessment

Falcon[63] addresses machine unlearning for removing harmful knowledge from LLMs, not activation steering for safety. The null-space projection in the original paper constrains steering transformations to preserve benign prompt activations, while Falcon[63] uses orthogonal gradient projection to resolve conflicts between forgetting and retention objectives during unlearning—fundamentally different technical approaches and application domains.

---

3. Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection

URL: View paper

Brief Assessment

Latent Space Obfuscation[61] focuses on privacy-preserving data obfuscation through latent space projection in autoencoders for medical and financial applications, not on activation steering or null-space constraints for LLM safety.

---

4. Robust Long-Term Vehicle Trajectory Prediction Using Link Projection and a Situation-Aware Transformer

URL: View paper

Brief Assessment

Link Projection Trajectory[64] addresses vehicle trajectory prediction on roads using geometric projection methods, not activation steering or utility preservation in language models. The domains are entirely different (autonomous driving vs. LLM safety).

---

5. DeCodec: Rethinking Audio Codecs as Universal Disentangled Representation Learners

URL: View paper

Brief Assessment

DeCodec[58] applies orthogonal projection to decompose audio representations into speech and background sound subspaces, not to preserve utility on benign prompts in LLM safety contexts. The technical domains and objectives are entirely different.

---

6. Learning a Generalizable Trajectory Sampling Distribution for Model Predictive Control

URL: View paper

Brief Assessment

Trajectory Sampling Distribution[59] focuses on learning sampling distributions for model predictive control in robotics navigation tasks, not on activation steering or null-space projection methods for preserving utility in language models.

---

7. Quantum-inspired Embeddings Projection and Similarity Metrics for Representation Learning

URL: View paper

Brief Assessment

Quantum Inspired Embeddings[57] focuses on quantum-inspired projection heads for embedding compression in representation learning, not on null-space projection methods for preserving utility in activation steering or LLM safety applications.

---

8. Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adjustment in Large Language Models

URL: View paper

Brief Assessment

Jailbreak Antidote[65] uses sparse representation adjustment (modifying ~5% of internal states) rather than null-space projection methods. The technical approaches differ fundamentally in their mathematical formulation and constraints.

---

9. BOF steelmaking endpoint carbon content and temperature soft sensor model based on supervised weighted local structure preserving projection

URL: View paper

Brief Assessment

BOF Steelmaking Sensor[60] addresses dimension reduction in industrial steelmaking processes using manifold learning techniques, not activation steering or utility preservation in language models. The technical domains are entirely distinct.

---

10. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models

URL: View paper

Brief Assessment

UniErase[62] applies null-space projection in the context of machine unlearning to preserve general knowledge when forgetting specific data, whereas the original paper uses it for activation steering to preserve utility on benign prompts during safety enhancement. These are fundamentally different application domains with distinct technical objectives.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] AlphaSteer: Learning Refusal Steering with Principled Null-Space Constraint View paper
- [1] On prompt-driven safeguarding for large language models View paper
- [2] Sorry-bench: Systematically evaluating large language model safety refusal View paper
- [3] Rule based rewards for language model safety View paper
- [4] Refusal tokens: A simple way to calibrate refusals in large language models View paper
- [5] Refusal in language models is mediated by a single direction View paper
- [6] Automating steering for safe multimodal large language models View paper
- [7] Refusing Safe Prompts for Multi-modal Large Language Models View paper
- [8] Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms View paper
- [9] Safety is Not Only About Refusal: Reasoning-Enhanced Fine-tuning for Interpretable LLM Safety View paper
- [10] A Minimalist Approach to LLM Reasoning: from Rejection Sampling to Reinforce View paper
- [11] Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning View paper
- [12] Backtracking improves generation safety View paper
- [13] Safety Layers in Aligned Large Language Models: The Key to LLM Security View paper
- [14] Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning View paper
- [15] Refusal-Aware Red Teaming: Exposing Inconsistency in Safety Evaluations View paper
- [16] Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment View paper
- [17] Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning View paper
- [18] From hard refusals to safe-completions: Toward output-centric safety training View paper
- [19] Refusal-trained llms are easily jailbroken as browser agents View paper
- [20] Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback View paper
- [21] Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training View paper
- [22] Safety pretraining: Toward the next generation of safe ai View paper
- [23] Safeconstellations: Steering llm safety to reduce over-refusals through task-specific trajectory View paper
- [24] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models View paper
- [25] Finding and Reactivating Post-Trained LLMs' Hidden Safety Mechanisms View paper
- [26] SafeSwitch: Steering Unsafe LLM Behavior via Internal Activation Signals View paper
- [27] When Safety Blocks Sense: Measuring Semantic Confusion in LLM Refusals View paper
- [28] From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law View paper
- [29] Controllable safety alignment: Inference-time adaptation to diverse safety requirements View paper
- [30] Steering without side effects: Improving post-deployment control of language models View paper
- [31] Llms encode harmfulness and refusal separately View paper
- [32] Internal activation as the polar star for steering unsafe llm behavior View paper
- [33] A Voter-Based Stochastic Rejection-Method Framework for Asymptotically Safe Language Model Outputs View paper
- [34] The devil behind the mask: An emergent safety vulnerability of diffusion llms View paper
- [35] Understanding and Mitigating Over-refusal for Large Language Models via Safety Representation View paper
- [36] Rule based rewards for fine-grained llm safety View paper

- [37] Understanding Refusal in Language Models with Sparse Autoencoders View paper
- [38] Should LLM Safety Be More Than Refusing Harmful Instructions? View paper
- [39] HumorReject: Decoupling LLM Safety from Refusal Prefix via A Little Humor View paper
- [40] COSMIC: Generalized Refusal Direction Identification in LLM Activations View paper
- [41] Robust LLM safeguarding via refusal feature adversarial training View paper
- [42] SafeLawBench: Towards Safe Alignment of Large Language Models View paper
- [43] Refusal Behavior in Large Language Models: A Nonlinear Perspective View paper
- [44] SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors View paper
- [45] Refusal Direction is Universal Across Safety-Aligned Languages View paper
- [46] Feature-Guided SAE Steering for Refusal-Rate Control using Contrasting Prompts View paper
- [47] Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary View paper
- [48] SARSteer: Safeguarding Large Audio Language Models via Safe-Ablated Refusal Steering View paper
- [49] Scaling laws for activation steering with Llama 2 models and refusal mechanisms View paper
- [50] LatentGuard: Controllable Latent Steering for Robust Refusal of Attacks and Reliable Response Generation View paper
- [51] DynaGuide: Steering Diffusion Polices with Active Dynamic Guidance View paper
- [52] DPD-LoRA: Dynamic Prompt-Driven Low-Rank Adaptation for Improved Generalization View paper
- [53] FairSteer: Inference Time Debiasing for LLMs with Dynamic Activation Steering View paper
- [54] Fusion Steering: Prompt-Specific Activation Control View paper
- [55] Understanding Reasoning Mechanisms in Large Language Models Through Direction Learning View paper
- [56] Steering LLMs' Reasoning With Activation State Machines View paper
- [57] Quantum-inspired Embeddings Projection and Similarity Metrics for Representation Learning View paper
- [58] DeCodec: Rethinking Audio Codecs as Universal Disentangled Representation Learners View paper
- [59] Learning a Generalizable Trajectory Sampling Distribution for Model Predictive Control View paper
- [60] BOF steelmaking endpoint carbon content and temperature soft sensor model based on supervised weighted local structure preserving projection View paper
- [61] Data Obfuscation Through Latent Space Projection for Privacy-Preserving AI Governance: Case Studies in Medical Diagnosis and Finance Fraud Detection View paper
- [62] UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models View paper
- [63] Falcon: Fine-grained activation manipulation by contrastive orthogonal unalignment for large language model View paper
- [64] Robust Long-Term Vehicle Trajectory Prediction Using Link Projection and a Situation-Aware Transformer View paper
- [65] Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adjustment in Large Language Models View paper
- [66] Privacy by Projection: Federated Population Density Estimation by Projecting on Random Features View paper
- [67] Pixel: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration View paper
- [68] Msrs: Adaptive multi-subspace representation steering for attribute alignment in large language models View paper
- [69] What makes and breaks safety fine-tuning? a mechanistic study View paper
- [70] MOSAICO: offline synthesis of adaptation strategy repertoires with flexible trade-offs View paper
- [71] MEUV: Achieving Fine-Grained Capability Activation in Large Language Models via Mutually Exclusive Unlock Vectors View paper
- [72] A Multi-Task Energy-Aware Impedance Controller for Enhanced Safety in Physical Human-Robot Interaction View paper