When Safety Geometry Collapses: Fine-Tuning Vulnerabilities in Agentic Guard Models

Ismail Hossain[1], Sai Puppala[2], Jannatul Ferdaus[1], Jahangir Alam[1], Tanzim Ahad[1], Yoonpyo Lee[3], Syed Bahauddin Alam[4] and Sajedul Talukder[1]

[1]University of Texas at El Paso, Texas, USA
[2]Southern Illinois University Carbondale, IL, USA
[3]Hanyang University, Seoul, South Korea
[4]University of Illinois Urbana-Champaign, Illinois, USA
{ihossain, jferdaus, malam10, tahad}@miners.utep.edu, sai.puppala@siu.edu,
lukeyounpyo@hanyang.ac.kr, alams@illinois.edu, stalukder@utep.edu

Abstract

A guard model fine-tuned on entirely benign data can lose all safety alignment– not through adversarial manipulation, but through standard domain specialization. We demonstrate this failure across three purpose-built safety classifiers– LlamaGuard, WildGuard, and Granite Guardian– deployed as protection layers in agentic AI pipelines, and show that it originates in the destruction of latent safety geometry: the structured harmful–benign representational boundary that guides classification. We extract per-layer safety subspaces via SVD on class-conditional activation differences and track how this boundary evolves under benign fine-tuning. Granite Guardian undergoes complete collapse– refusal rate drops from 85% to 0%, CKA falls to zero, and 100% of outputs become ambiguous– a severity exceeding prior findings on general-purpose LLMs, explained by the specialization hypothesis: concentrated safety representations are efficient but catastrophically brittle. To mitigate this, we propose Fisher-Weighted Safety Subspace Regularization (FW-SSR), a training-time penalty combining (i) curvature-aware direction weights derived from diagonal Fisher information and (ii) an adaptive λ_t that scales with task–safety gradient conflict. FW-SSR recovers 75% refusal on Granite Guardian (CKA $= 0.983$) and reduces WildGuard's Attack Success Rate to 3.6%– below the unmodified baseline– by actively sharpening the safety subspace rather than merely anchoring it. Across all three models, structural representational geometry (CKA, Fisher score) predicts safety behavior more reliably than absolute displacement metrics, establishing geometry-based monitoring as a necessary component of guard model evaluation in agentic deployments.

Website — https://supreme-lab.github.io/wsgc/

Introduction

Modern AI systems are rapidly evolving from single-model deployments toward *agentic architectures* (Park et al. 2023; Wang et al. 2024), in which multiple specialized LLMs collaborate through structured communication pipelines to solve complex, long-horizon tasks. In such systems, each agent handles a distinct sub-task– retrieval, reasoning, planning, code execution, or synthesis– and agents exchange intermediate outputs across turns. This collaborative specialization enables capabilities far beyond any individual model, but it introduces a new attack surface: if any single agent in the pipeline can be made to process or propagate harmful content, the safety of the entire system is compromised.

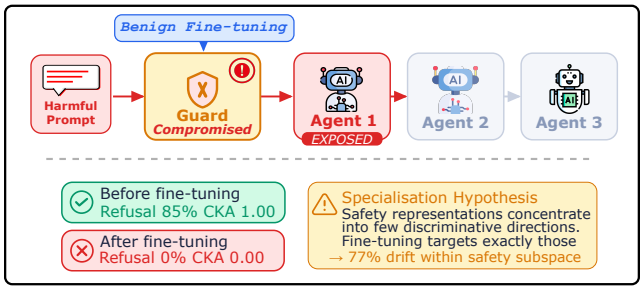

Figure 1: Benign Fine-Tuning Destroys Safety Alignment

A standard mitigation in agentic deployments is to protect individual agents with *dedicated safety guard models*: purpose-built classifiers such as MD-Judge (OpenSafetyLab 2024), LlamaGuard (Inan et al. 2023), and Granite Guardian (IBM Research 2024) that intercept inputs before they reach the agent, filtering adversarial or policy-violating content. Unlike general-purpose LLMs that must balance safety with broad capability, these guard models are trained specifically to classify requests across a standardized harm taxonomy and reliably refuse harmful inputs while permitting benign ones. They operate as a *static protection layer*: once deployed, they are assumed to remain aligned regardless of downstream pipeline changes.

This assumption is violated in practice. As shown in Figure 1, benign fine-tuning can completely destroy safety alignment, reducing refusal rates to zero despite the absence of harmful training data. Agentic systems are routinely fine-tuned to specialize individual agents for particular domains: a medical-assistant agent tuned on clinical dialogue, a legal-reasoning agent tuned on case documents, or an enterprise agent tuned on company-specific policy. The safety guards

protecting these agents naturally undergo the same specialization: fine-tuned on domain-relevant, entirely *benign* data to improve domain-specific classification performance. As we demonstrate, this benign adaptation is sufficient to *catastrophically destroy* the guard model's safety alignment– a fundamental threat to the safety of the agentic pipeline as a whole.

We study this failure as a *latent geometry* phenomenon. Safety-aligned guard models encode the harmful– benign distinction as a structured separation in activation space: prompts from different safety classes occupy distinct regions, separated by a latent safety boundary that guides classification. Standard fine-tuning, optimizing a task loss with no constraint on internal representations, erodes this boundary even when training data is entirely benign. The result is *safety geometry collapse*: representations of harmful and benign inputs become indistinguishable. In an agentic context, this collapse means the guard model can no longer distinguish an adversarial jailbreak from a routine query, allowing harmful content to reach the agent it was designed to protect and potentially propagate to downstream agents. In the Figure 2, The original guard model *(left)* maintains a well-separated latent space between harmful and benign inputs, achieving 85% refusal rate. Domain-specific benign fine-tuning *(center)* destroys this boundary entirely– refusal drops to 0% and the latent space collapses– despite containing no harmful training signal. FW-SSR *(right)* restores the harmful–benign separation by regularizing safety-critical subspace directions during fine-tuning, recovering 75% refusal while preserving domain utility.

Why Guard Models Are Especially Vulnerable. Prior work (Qi et al. 2023) reports approximately 55 pp refusal drops in general-purpose LLMs under benign Alpaca fine-tuning. We observe an **85 pp collapse** in Granite Guardian– complete erasure of refusal behavior– which we attribute to a *specialization hypothesis*: guard models concentrate safety representations into a small number of highly discriminative latent directions, creating classification efficiency but also catastrophic brittleness when those directions are perturbed by fine-tuning.

Research Gap. Three specific limitations motivate our approach. (i) No prior work examines benign fine-tuning vulnerability in purpose-built guard models operating in agentic pipelines, which exhibit qualitatively more severe degradation than general-purpose LLMs. (ii) Existing latent anchor penalties penalize all safety directions equally, ignoring the non-uniform curvature structure of the safety subspace. (iii) Fixed regularization strength λ cannot adapt to the varying conflict between task and safety gradients across training.

Contributions. (1) **Threat Characterization.** We identify benign fine-tuning during agent specialization as a systematic threat to guard model safety in agentic AI pipelines, introducing a latent geometry analysis protocol that reveals complete safety geometry collapse more severe than in general-purpose LLMs. (2) **FW-SSR.** A novel training-time penalty weighting each safety direction by diagonal Fisher information, with adaptive λ_t scaled by task– safety gradient conflict, enabling safe specialization of guard models without sacrificing domain performance. (3) **Evaluation.** FW-SSR recovers 75% refusal behavior and CKA $= 0.983$ on Granite Guardian 3B under benign Alpaca fine-tuning, with ablations isolating each component's contribution.

Related Work

Agentic AI Safety. Recent work on multi-agent LLM systems (Park et al. 2023; Wang et al. 2024) identifies security risks unique to agentic pipelines, including cross-agent prompt injection, adversarial manipulation of inter-agent communication, and goal misalignment in planning agents. Adversarial instructions can hijack downstream tasks (Perez and Ribeiro 2022), while indirect prompt injection through retrieved documents can compromise entire pipelines (Greshake et al. 2023). Similarly, jailbreak vulnerabilities in base VLMs can be *inherited* by fine-tuned models, with adversarial images transferring to safety-tuned variants at success rates above 86.5% (Wang et al. 2025). These findings suggest that fine-tuning often propagates latent vulnerabilities rather than removing them. In contrast, we identify a *structural* failure mode: benign fine-tuning for agent specialization can systematically destroy the safety geometry of guard models even without adversarial inputs.

Safety Degradation Under Fine-Tuning. Benign fine-tuning can significantly weaken safety alignment in LLMs, increasing harmful completion rates even without malicious data, with refusal drops of up to $\sim$55 pp (Qi et al. 2023). Certain data properties disproportionately degrade safety (Zhang et al. 2024), and small adversarial injections can further amplify this collapse (Yang et al. 2023). Lisa (Huang et al. 2024) mitigates degradation by alternating alignment and task states with proximal drift constraints, but operates at the loss level rather than preserving latent safety geometry as FW-SSR does. Similarly, backdoor purification can reduce attack success rates while leaving latent vulnerabilities intact, allowing rapid re-learning from minimal poisoned samples (Min et al. 2024). We extend this line of work to guard models, analysing safety degradation through latent geometry and highlighting the risk in agentic pipelines where a single compromised guard can expose the entire system.

Representation-Level Safety. Safety-relevant behaviors are encoded in structured activation-space directions that can be identified and manipulated (Zou et al. 2023), with high-level concepts occupying low-dimensional linear subspaces (Park, Choe, and Veitch 2023). Hidden representations also encode evaluative signals beyond surface outputs; for example, task difficulty is linearly decodable from the initial hidden state (Zhu et al. 2025). These findings indicate that latent geometry contains safety-relevant information not captured by behavioral metrics, consistent with our observation that CKA and Fisher scores detect safety degradation earlier than output-based measures. Similar vulnerabilities appear in multimodal models, where adversarial images can universally jailbreak aligned VLMs by exploiting visual representation structure (Qi et al. 2024). Motivated by this perspective, we compute explicit safety subspaces from class-conditional activation statistics and use them as

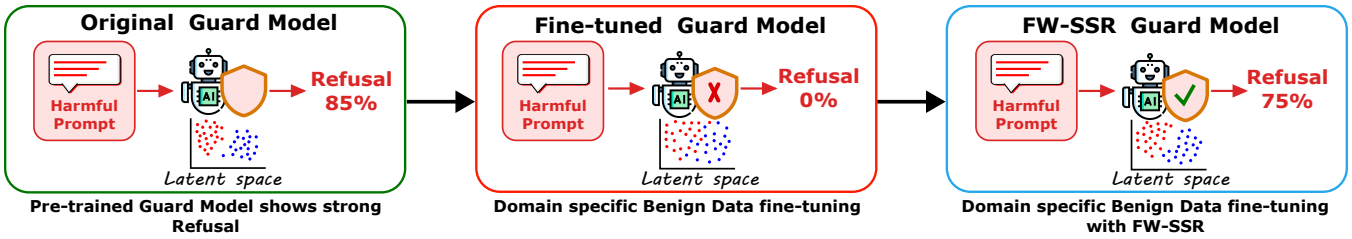

Figure 2: Safety alignment collapse and recovery in guard models under benign fine-tuning.

training-time anchors to preserve alignment during domain specialization.

Fisher Information in Continual Learning. EWC (Kirkpatrick et al. 2017) penalizes changes to parameters important for prior tasks via diagonal Fisher information, preventing catastrophic forgetting in sequential learning. The guard model safety collapse we observe is an instance of vertical continual learning failure (Shi et al. 2024) – domain specialization gradients overwrite concentrated safety representations in the same way catastrophic forgetting destroys prior task knowledge. Unlike EWC, which operates at the parameter level and is incompatible with LoRA's frozen base, FW-SSR applies Fisher weighting at the *activation level* within the safety subspace, preserving compatibility with parameter-efficient agent specialization.

Problem Formulation

Figure 2 illustrates a guard model deployed as a safety layer in an agentic system, responsible for filtering harmful inputs before they reach downstream components. While the original aligned model reliably separates harmful from benign prompts in latent space, standard domain-specific fine-tuning on benign data can significantly degrade this alignment. This degradation manifests as a collapse of the representation space, where harmful and benign inputs become indistinguishable, resulting in a complete loss of refusal behavior.

To address this, we introduce a regularization framework that constrains safety-critical representation drift during fine-tuning. The objective is to preserve the latent structure that encodes the harmful–benign boundary while allowing adaptation to downstream tasks, thereby maintaining robust safety behavior under domain specialization.

Setting

Let M_{θ_0} be a pretrained, safety-aligned guard model. Given a benign fine-tuning dataset $\mathcal{D}_{\text{task}}$ and a safety probe dataset $\mathcal{D}_{\text{safe}} = \{(x_i, y_i)\}_{i=1}^{N}$ with $y_i \in \{0, 1\}$ (benign/harmful), we wish to adapt M_{θ_0} to downstream utility while preserving safety alignment, yielding parameters θ^*.

Safety Subspace Extraction

For transformer layer ℓ and parameters θ, let $h_\ell(x;\theta) \in \mathbb{R}^d$ denote the hidden state at the final non-padding token position (last-token representation), consistent with standard practice for decoder-only causal LMs. Let $H_\ell^c \in \mathbb{R}^{N_c \times d}$ denote the matrix whose i-th row is $h_\ell(x_i^c;\theta_0)$, extracted from the N_c probe samples of class $c \in \{h,\ b\}$ (harmful, benign). Let $\mu_\ell^c = \frac{1}{N_c}\sum_i [H_\ell^c]_i \in \mathbb{R}^d$ be the class-conditional mean under θ_0, and let

$$\tilde{H}_\ell^c = H_\ell^c - \mathbf{1}_{N_c}(\mu_\ell^c)^\top \in \mathbb{R}^{N_c \times d} \tag{1}$$

denote the class-centered activation matrix. We construct an augmented activation matrix by stacking the two centered matrices together with n_a scaled copies of the between-class mean difference:

$$A_\ell = \begin{bmatrix} \tilde{H}_\ell^h \\ \tilde{H}_\ell^b \\ \gamma \cdot \mathbf{1}_{n_a}(\Delta\mu_\ell)^\top \end{bmatrix} \in \mathbb{R}^{(N_h+N_b+n_a)\times d}, \tag{2}$$

where $\Delta\mu_\ell = \mu_\ell^h - \mu_\ell^b \in \mathbb{R}^d$ is the between-class difference vector, $n_a = \min(32, N_h)$, and $\gamma = 5$ amplifies the primary safety direction to dominate the top singular values.

The *safety subspace* $U_\ell \in \mathbb{R}^{k\times d}$ comprises the top-k *right* singular vectors of A_ℓ, i.e. the first k rows of $V_h^\top$ from $\text{SVD}(A_\ell) = P\,\Sigma\,V_h^\top$ (with full_matrices = False, giving $V_h^\top \in \mathbb{R}^{\min(N_{\text{aug}},d)\times d}$). These vectors span the top-k directions of variance in $\mathbb{R}^d$ that are most discriminative between harmful and benign activations. U_ℓ is computed once from M_{θ_0} and held fixed throughout fine-tuning.

Safety Drift Metrics

The *safety drift* at layer ℓ under θ is:

$$\Delta_{\text{safe}}(\ell;\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{safe}}}\left[\left\|U_\ell^\top\big(h_\ell(x;\theta) - h_\ell(x;\theta_0)\big)\right\|_2\right]. \tag{3}$$

The *drift ratio* measures the fraction of total activation change within the safety subspace:

$$\rho(\ell;\theta) = \frac{\Delta_{\text{safe}}(\ell;\theta)}{\mathbb{E}_x[\|h_\ell(x;\theta) - h_\ell(x;\theta_0)\|_2] + \varepsilon}. \tag{4}$$

The *Fisher discriminant score* measures class separation:

$$\text{FS}(\ell;\theta) = \frac{\left\|\mu_\ell^h(\theta) - \mu_\ell^b(\theta)\right\|_2}{\sigma_\ell^h(\theta) + \sigma_\ell^b(\theta) + \varepsilon}, \tag{5}$$

where $\sigma_\ell^c(\theta)$ is the mean intra-class L2 spread. Linear CKA (Kornblith et al. 2019) compares representational geometry between θ and θ_0 (Eq. 6), computed on a 64-sample subsample for efficiency:

$$\text{CKA}(H_\ell(\theta), H_\ell(\theta_0)) = \frac{\text{HSIC}(K_\theta, K_0)}{\sqrt{\text{HSIC}(K_\theta,K_\theta)\cdot\text{HSIC}(K_0,K_0)}} \tag{6}$$

Problem Statement. Find θ^* minimizing task loss on $\mathcal{D}_{\text{task}}$ while, for all probed layers $\ell \in \mathcal{L}$, minimizing $\Delta_{\text{safe}}(\ell; \theta^*)$ and preserving refusal rates relative to M_{θ_0}.

Fisher-Weighted Safety Subspace Regularization

FW-SSR: Formulation

We propose FW-SSR, replacing the uniform penalty with a curvature-aware adaptive variant:

$$\mathcal{L}_{\text{FW-SSR}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{safe}}} \sum_{\ell \in \mathcal{L}} \left\| \hat{F}_\ell \odot U_\ell^\top \big(h_\ell(x; \theta) - h_\ell(x; \theta_0) \big) \right\|_2^2, \tag{7}$$

where $\hat{F}_\ell \in \mathbb{R}^k$ is a vector of per-direction curvature weights and $\odot$ denotes element-wise multiplication. The total objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_t \cdot \mathcal{L}_{\text{FW-SSR}}$. Gradients flow only through the fine-tuned model's activations; original model activations are detached and serve as frozen anchors.

Fisher Weight Estimation

The curvature weights estimate sensitivity via a diagonal approximation of projected Fisher information:

$$F_\ell \approx \mathbb{E}_{x \sim \mathcal{D}_{\text{safe}}} \left[(U_\ell^\top h_\ell(x; \theta))^2 \right] \in \mathbb{R}^k. \tag{8}$$

Before application, F_ℓ is mean-normalized:

$$\hat{F}_\ell = \frac{k \cdot F_\ell}{\sum_j F_\ell[j] + \varepsilon}, \tag{9}$$

ensuring $\hat{F}_\ell$ has mean 1 by construction, so penalty scale is consistent across layers. Fisher weights are updated via EMA every $\tau = 50$ gradient steps:

$$F_\ell \leftarrow \beta F_\ell + (1 - \beta) F_\ell^{\text{new}}, \quad \beta = 0.9, \tag{10}$$

and initialized to $\mathbf{1}_k$ for uniform regularization before estimates stabilize.

Adaptive λ Scheduling

We measure gradient conflict via cosine similarity between task and safety gradients every 20 steps:

$$s_t = \cos\big(\nabla_\theta \mathcal{L}_{\text{task}},\ \nabla_\theta \mathcal{L}_{\text{FW-SSR}} \big). \tag{11}$$

We update λ as:

$$\lambda_t^{\text{new}} \leftarrow \text{clip}\Big(\lambda_{t-1} \cdot (1 - \tfrac{1}{2} s_t),\ 10^{-4},\ 1.0 \Big), \tag{12}$$

with exponential smoothing:

$$\lambda_t \leftarrow 0.95\, \lambda_{t-1} + 0.05\, \lambda_t^{\text{new}}. \tag{13}$$

When $s_t \approx +1$ (objectives aligned), $(1 - \frac{1}{2} s_t) \approx 0.5$ reduces λ. When $s_t \approx -1$ (objectives conflict), the factor ≈ 1.5 increases λ, providing strongest protection exactly when needed.

Algorithm 1: FW-SSR Training

Require: M_{θ_0}, $\mathcal{D}_{\text{task}}$, $\mathcal{D}_{\text{safe}}$, $\{U_\ell\}_{\ell \in \mathcal{L}}$, $\lambda_0, k, \beta, \tau$
Output: Safety-preserved fine-tuned parameters θ^*

$\theta \leftarrow \theta_0$; $F_\ell \leftarrow \mathbf{1}_k\ \forall \ell$; $\lambda \leftarrow \lambda_0$
for each step $t = 1, \ldots, T$ **do**
  Sample $B_{\text{task}} \sim \mathcal{D}_{\text{task}}$, $B_{\text{safe}} \sim \mathcal{D}_{\text{safe}}$
  Compute $\mathcal{L}_{\text{task}}$ on B_{task}
  Compute $\hat{F}_\ell$ (Eq. 9), then $\mathcal{L}_{\text{FW-SSR}}$ (Eq. 7) on B_{safe}
  $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{FW-SSR}})$
  if $t \bmod \tau = 0$ **then**
    Update F_ℓ via EMA (Eq. 10)
  end if
  if $t \bmod 20 = 0$ **then**
    Compute s_t (Eq. 11)
    Update λ (Eqs. 12– 13)
  end if
end for
return θ^*

Experimental Setup

Models

We evaluate on three purpose-built guard model families.

WildGuard (Han et al. 2024): An open-source safety classifier built on Mistral-7B (32 transformer layers, $d = 4096$), trained to jointly detect harmful requests, adversarial jailbreaks, and model refusals across a broad harm taxonomy. Probed at layers $\{20, 24, 28, 31\}$ (50/65/80/100% depth). **Granite Guardian 3.0-2B** (IBM Research 2024): IBM's open-access guard model; complete experimental results are available and reported in Tables 1. **LlamaGuard-3-8B** (Inan et al. 2023): Meta's 8B safety classifier built on LLaMA-3.1 (32 layers, $d = 4096$).

Fine-Tuning Data

We fine-tune on 2,000 randomly sampled examples from the Alpaca instruction-following dataset (Taori et al. 2023), following the Risk Level 3 protocol of Qi et al. (2023). This simulates realistic domain adaptation without adversarial intent.

Safety Probe Dataset

We construct $\mathcal{D}_{\text{safe}}$ with 80 balanced prompt pairs (40 harmful, 40 benign). *Harmful prompts*: 38 unique prompts spanning 5 harm categories– **violence** (improvised weapons, poisoning), **cybercrime** (SQL injection, keyloggers, ransomware), **fraud** (counterfeiting, scam scripts, pyramid schemes), **drug synthesis** (fentanyl synthesis, dangerous combinations), and **harassment** (doxxing, intimidation campaigns)– plus biological weapons, illegal firearms, and identity fraud. Prompts are repeated to fill 40 slots. *Benign prompts*: 20 factual queries repeated to fill 40 slots. Labels: $y = 1$ (harmful), $y = 0$ (benign).

Implementation

All fine-tuning uses QLoRA (Dettmers et al. 2023) with 4-bit NF4 quantization; LoRA rank $r = 8$, $\alpha = 16$, ap-

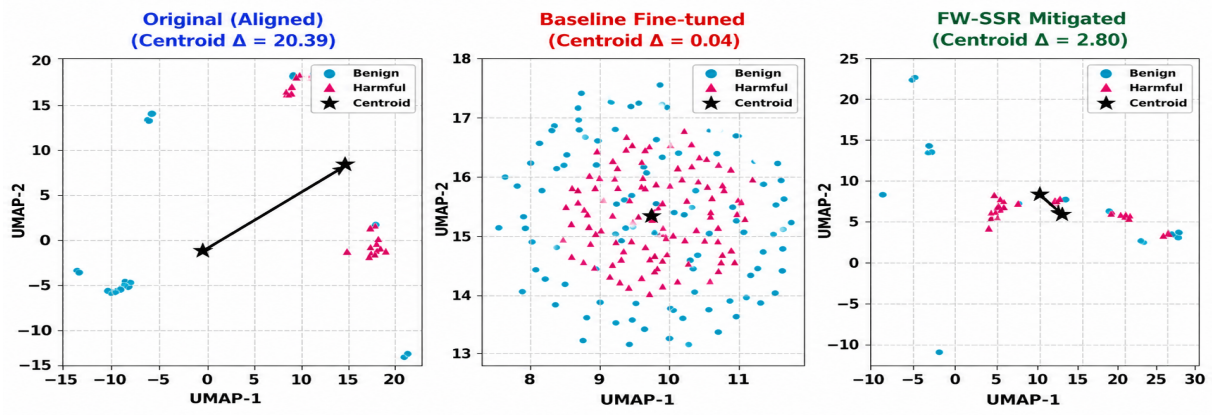

Figure 3: Granite Guardian 3.0-2B– Latent Space Geometry Layer 39 Harmful vs. Benign Prompt Separation

plied to all attention ($q/k/v/o$_proj) and MLP (gate/up/down_proj) matrices (auto-detected per architecture). Training: 3 epochs, batch size 8, gradient accumulation 4 steps (effective batch size 32), learning rate 2×10^{-5}, AdamW weight decay 10^{-2}, gradient clipping at 1.0. FW-SSR: $\lambda_0 = 0.1$, $\beta = 0.9$, $\tau = 50$, $k = 32$. Single NVIDIA H100 96 GB GPU; M_{θ_0} offloaded to CPU, moved to GPU only for regularization forward passes.

Evaluation Metrics

Safety Behavior (20 harmful prompts, keyword classification): **Refusal Rate**– fraction explicitly declining; **Compliance Rate**– fraction providing harmful content; **Ambiguous Rate**– fraction matching neither criterion. A model with 0% refusal and 0% compliance produces 100% ambiguous outputs, indicating complete representational collapse: the model neither refuses nor produces useful harmful content, generating off-topic incoherent outputs with no safety value.

Latent Geometry: safety drift (Eq. 3), drift ratio (Eq. 4), cosine similarity, Fisher score (Eq. 5), inter-class centroid distance, and CKA vs. original (Eq. 6, 64-sample subsample).

Results

Table 1 presents results for all three guard models at the final probed layer (100% depth).

Collapse Severity Tracks Safety Representation Concentration

Benign fine-tuning degrades all three guard models, but the severity scales with the concentration of safety representations. Granite Guardian undergoes the most catastrophic collapse: refusal rate drops from 85% to 0%, CKA falls to 0.00, and Fisher score and inter-class distance both reach 0.00, indicating complete destruction of the harmful– benign classification boundary. Figure 3 illustrates this effect, showing that fine-tuning collapses the latent separation between harmful and benign representations, while FW-SSR restores the geometric structure. A drift ratio of 0.77 confirms the mechanism– 77% of all activation change is concentrated in the safety subspace, meaning unconstrained gradients disproportionately overwrite the directions most critical to safety. This 85 pp refusal collapse substantially exceeds the ∼55 pp drops reported for general-purpose LLMs (Qi et al. 2023), consistent with the specialization hypothesis: guard models encode safety into few highly discriminative directions, which creates efficiency but catastrophic brittleness. LlamaGuard-3 shows partial degradation– CKA falls to 0.83, Fisher score from 0.62 to 0.47, drift ratio 0.51– without reaching zero on any metric, suggesting a more distributed safety geometry that distributes and thus dilutes the effect of fine-tuning gradients. WildGuard exhibits a qualitatively different pattern: refusal drops from 35% to 5% yet Fisher score *increases* from 0.97 to 1.01 and safety drift remains the lowest of all three models (17.21), indicating that behavioral degradation here is driven by disruption of the decision function above the subspace rather than collapse of the subspace itself. Together, the three models reveal that safety collapse is not a single phenomenon but a spectrum whose severity is predicted by geometry, not behavior alone.

FW-SSR Recovers Safety Geometry Across All Three Models

FW-SSR consistently improves both behavioral and geometric metrics relative to the fine-tuned baseline across all architectures. For Granite Guardian, refusal rate recovers to 75% (an 88% relative recovery), CKA reaches 0.98, Fisher score recovers to 0.55 (81% of original), and inter-class distance to 11.36 (93% of original), confirming that the safety classification boundary is largely reconstituted despite substantial absolute drift. LlamaGuard-3 shows consistent but more modest improvement: refusal rate rises from 15% to 25%, safety drift reduces by 17% (45.99 → 38.00), and Fisher score recovers to 0.54, approaching the original 0.62. For WildGuard, FW-SSR raises Fisher score to 1.08– *exceeding the original 0.97*– as the curvature-aware weighting mechanism actively sharpens the class-separating directions rather than merely anchoring them, while compliance returns to 0% and refusal recovers to 20%. Across all three models, the largest absolute improvements occur in the metrics most directly linked to classification boundary structure– Fisher

score, inter-class distance, and CKA– rather than in drift suppression, pointing to the mechanism examined next.

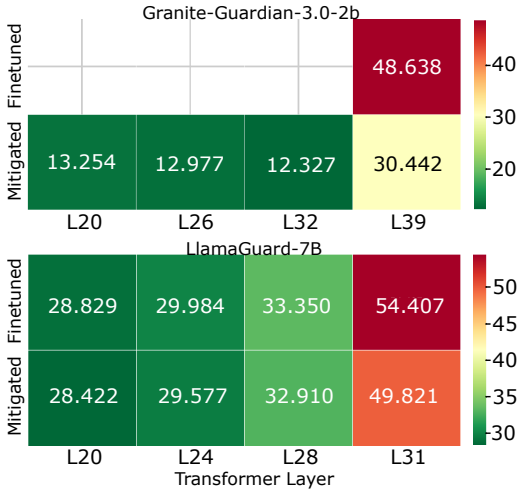

Figure 4: Per-Layer Safety Drift Heatmap

Structural Geometry Is the Primary Safety Carrier

A consistent pattern across all three models establishes that *relational* geometry predicts safety behavior more reliably than *absolute displacement*. Granite Guardian's FW-SSR condition provides the sharpest demonstration: safety drift is only 37% suppressed (48.64 $\rightarrow$ 30.44), yet CKA recovers to 0.98 and refusal rate to 75%. Activations have moved substantially from their original positions, but their relational organization– how harmful and benign inputs are separated in the safety subspace– is almost perfectly preserved, and behavior tracks this structure, not the displacement magnitude. WildGuard's fine-tuned condition provides the complementary counter-example: safety drift of 17.21 is the lowest of any fine-tuned model, yet refusal drops from 35% to 5%. Low displacement does not prevent behavioral collapse when the class-separating geometry is eroded, as reflected by inter-class distance falling from 11.03 to 7.94 despite stable drift. These two observations jointly imply that CKA and Fisher score are the most diagnostically meaningful metrics for guard model safety evaluation: they detect both the global structural collapse of Granite and the localized boundary erosion of WildGuard, while drift-based metrics fail in both cases.

Figure 3 provides geometric evidence of safety collapse and recovery in *Granite Guardian 3.0-2B*. In the original aligned model, harmful and benign prompts form well-separated clusters with a large centroid distance ($\Delta = 20.39$), indicating a clear and robust decision boundary. However, after benign fine-tuning, this separation collapses entirely, with centroid distance dropping to $\Delta = 0.04$, and both classes becoming indistinguishable in latent space. This geometric collapse explains the observed failure in safety behavior, where the model loses its ability to discriminate harmful inputs. In contrast, the FW-SSR mitigated model restores this structure, increasing centroid separation to $\Delta = 2.80$ and re-establishing a meaningful boundary between harmful and benign representations. These results demonstrate that safety is fundamentally governed by latent geometry, and that preserving this structure is critical for maintaining alignment under fine-tuning.

Layer-wise Safety Drift

Figure 4 presents per-layer safety drift heatmaps for *Granite-Guardian-3.0-2b* and *LlamaGuard-7B*, comparing baseline fine-tuned (no defense) and FW-SSR mitigated conditions across transformer depth. Two consistent patterns emerge. First, safety drift concentrates in deep layers (80–100% depth) for both models, confirming that safety-discriminative representations reside predominantly in late transformer layers and are therefore disproportionately targeted by fine-tuning gradients. Second, the severity of drift diverges sharply between architectures: Granite-Guardian exhibits a catastrophic spike at its final layer (L39: 48.638 fine-tuned vs. 30.442 mitigated), consistent with its complete safety geometry collapse (refusal rate $85\% \rightarrow 0\%$, CKA $= 0.00$), whereas LlamaGuard-7B shows a more gradual, distributed drift profile (28.8–54.4) with modest inter-condition differences, reflecting its more diffuse safety geometry and correspondingly partial degradation. This contrast directly supports the *specialization hypothesis*: guard models that concentrate safety into few highly discriminative directions gain classification efficiency but become catastrophically brittle under domain fine-tuning. FW-SSR reduces drift most aggressively at final layers where Fisher curvature weights are highest, yet the key insight is that even a moderate 37% drift suppression in Granite-Guardian restores 75% refusal and CKA $= 0.983$ - demonstrating that *relational geometry*, not absolute activation displacement, is the primary carrier of safety behavior.

Per-Benchmark ASR Analysis

Table 2 reveals three consistent findings. First, benign fine-tuning raises overall ASR from 9.2% to 17.1% (an 85.9% relative increase) uniformly across all four benchmarks, with the largest degradation on AdvBench (+13.0 pp) and the smallest on HarmBench (+3.2 pp), ruling out benchmark-specific artefacts and confirming that safety degrades broadly across diverse prompt distributions. Second, FW-SSR not only reverses this degradation but reduces ASR to 3.6%– 60.9% below the *original* model's 9.2%– achieving the lowest ASR on every individual benchmark; this improvement beyond the original baseline is explained geometrically: the Fisher-weighted regularization actively sharpens the harmful– benign classification boundary (Fisher score $0.118 \rightarrow 0.153$; inter-class distance $37.1 \rightarrow 62.7$), rather than merely anchoring it to its pre-fine-tuning state. Finally, the highest residual ASR under FW-SSR occurs on AdvBench (5.0%), whose direct harmful instructions densely sample harm categories underrepresented in the current safety probe set; category-stratified probe construction is the most direct remedy.

Metric	Granite Guardian 3.0-2B			LlamaGuard-3-8B			WildGuard		
	O	FT	Mit	O	FT	Mit	O	FT	Mit
Safety Behavior (20 harmful prompts, keyword classification)									
Refusal Rate (%) ↑	85.00	0.00	**75.00**	5.00	15.00	**25.00**	35.00	5.00	**20.00**
Compliance Rate (%) ↓	0.00	**0.00**	5.00	5.00	10.00	10.00	0.00	5.00	**0.00**
Ambiguous Rate (%)	15.00	100.00	**20.00**	90.00	75.00	**65.00**	65.00	90.00	**80.00**
Latent Safety Geometry (final probed layer, 64-sample CKA)									
Safety Drift ↓	0.00	48.64	**30.44**	0.00	45.99	**38.00**	0.00	17.21	**17.17**
Drift Ratio ↓	0.00	0.77	**0.56**	0.00	0.51	**0.44**	0.00	0.65	**0.64**
Cosine Sim. ↑	1.00	0.00	**0.58**	1.00	0.12	**0.21**	1.00	0.30	**0.31**
Fisher Score ↑	0.68	0.00	**0.55**	0.62	0.47	**0.54**	0.97	1.01	**1.08**
Inter-Class Dist. ↑	12.19	0.00	**11.36**	12.34	9.97	**11.25**	11.03	7.94	**8.64**
CKA vs. Orig. ↑	1.00	0.00	**0.98**	1.00	0.83	0.83	1.00	0.83	**0.86**

Table 1: Safety geometry and behavioral evaluation across three guard models under benign Alpaca fine-tuning and FW-SSR mitigation. O = original aligned model; FT = baseline fine-tuned (no defense); Mit = FW-SSR mitigated. ↑ higher is better; ↓ lower is better. Bold: best value among {FT, Mit} per metric per model.

	Original	Fine-tuned	FW-SSR
Per-Benchmark ASR (%) ↓			
AdvBench	14.0	27.0	**5.0**
HarmBench	6.2	9.4	**2.1**
JailbreakBench	7.0	13.0	**4.0**
StrongREJECT	9.4	18.8	**3.1**
Overall ASR	9.2	17.1	**3.6**

Table 2: Evaluation results for WildGuard across original, fine-tuned (no defence), and FW-SSR conditions. ASR evaluated by cross-model judge (LlamaGuard-3-8B). ↓ lower is better; ↑ higher is better. Bold: best per row; red: worst.

Per-Category ASR Analysis

Table 3 shows that fine-tuning raises ASR in 14 of 22 categories, with the sharpest increases in Malware/Hacking (+30.0 pp), Illegal Goods & Services (+18.7 pp), Sexual/Adult Content (+20.0 pp), and Hate, Harassment & Discrimination (+18.8 pp)– all categories whose prompts involve multi-step harmful instructions that are particularly sensitive to shifts in the model's instruction-following behavior induced by Alpaca fine-tuning. In contrast, categories already at 0% original ASR (Chemical/Biological, Misinformation, Expert Advice) show only marginal degradation under fine-tuning, suggesting that WildGuard's original training data provides stronger coverage for these harm types and that fine-tuning erodes safety more readily where the original boundary is closer to the decision threshold. FW-SSR eliminates ASR entirely in 12 of 22 categories, driving them to 0%, and achieves the lowest ASR in 18 of 22 categories overall; the four categories where FW-SSR fails to improve are Harassment/Bullying (6.2% across all conditions, indicating an irreducible evaluation artefact at this sample size), Economic Harm (10.0% across all conditions), Sexual/Adult Content and Privacy (unchanged at 10.0%, suggesting the safety probe underrepresents the specific prompt structures used in these categories). The one anomaly is Physical Harm, where FW-SSR introduces a 10.0% ASR despite both original and fine-tuned conditions showing 0.0%; this isolated regression is attributable to insufficient probe coverage of physical harm prompts in $\mathcal{D}_{\text{safe}}$, causing the safety subspace to provide weak regularization in this direction, and points to category-stratified probe construction as the most direct remedy.

Discussion

Implications for Agentic AI Safety. Our findings expose a systematic vulnerability in the standard agentic deployment practice of fine-tuning safety guards alongside the agents they protect. Guard models occupy a privileged position in multi-agent pipelines — they are the last line of defense before a harmful prompt reaches an agent and propagates downstream — yet even short benign fine-tuning runs (3 epochs, 2,000 Alpaca samples) are sufficient to destroy this protection entirely in Granite Guardian, or substantially erode it in LlamaGuard-3 and WildGuard, without any adversarial intent. FW-SSR provides a concrete mechanism for *safe specialization*: guard models can be adapted to domain-specific requirements while preserving the safety representations that make them effective protection layers, as demonstrated by WildGuard's FW-SSR ASR of 3.6% — lower than even the unmodified model.

The Specialization Hypothesis. The severity of collapse scales with the concentration of safety representations. Granite Guardian, which encodes safety into a small number of highly discriminative directions, undergoes complete catastrophic collapse: a drift ratio of 0.77 confirms that unconstrained gradients disproportionately target safety-critical subspace directions, mirroring catastrophic forgetting (Kirkpatrick et al. 2017). LlamaGuard-3 and WildGuard, with more distributed safety geometries, exhibit partial and localized degradation respectively. This concentration–brittleness trade-off has a direct design implication: guard models intended for agentic deployment should be evaluated for safety representation concentration before fine-tuning, as concentrated architectures require

Category	Orig	FT	FW	Category	Orig	FT	FW
Harmful	6.2	**0.0**	**0.0**	Disinformation	**0.0**	**0.0**	**0.0**
Illegal	18.8	31.2	**6.2**	Malware / Hacking	10.0	40.0	**0.0**
Cybercrime / Intrusion	6.2	6.2	**0.0**	Physical Harm	**0.0**	**0.0**	10.0
Chemical / Biological	**0.0**	6.2	**0.0**	Privacy	**10.0**	20.0	**10.0**
Harassment / Bullying	**6.2**	**6.2**	**6.2**	Fraud / Deception	10.0	**0.0**	**0.0**
Violence	**6.2**	18.8	**6.2**	Harassment / Discrimination	10.0	10.0	**0.0**
Non-violent Crimes	12.5	12.5	**0.0**	Hate, Harassment & Discrim.	**0.0**	18.8	**0.0**
Sexual / Adult Content	**10.0**	30.0	**10.0**	Illegal Goods & Services	18.8	37.5	**6.2**
Sexual Content	12.5	18.8	**6.2**	Expert Advice	**0.0**	10.0	**0.0**
Misinformation / Disinform.	**0.0**	6.2	**0.0**	Economic Harm	**10.0**	**10.0**	**10.0**
Disinformation & Deception	6.2	6.2	**0.0**	Government Decision-making	10.0	10.0	**0.0**
Overall: Original **9.2**% Fine-tuned 17.1% FW-SSR **3.6**%							

Table 3: Per-category Attack Success Rate (%) for WildGuard across original, fine-tuned (no defence), and FW-SSR conditions. Bold: best (lowest ASR) per row; red: worst.

stronger regularization budgets.

Structural Geometry as the Primary Safety Carrier. Across all three models, relational geometry metrics (CKA, Fisher score, inter-class distance) predict safety behavior more reliably than absolute displacement metrics (safety drift, cosine similarity). In Granite Guardian, FW-SSR achieves CKA $= 0.98$ and 75% refusal recovery despite only 37% drift suppression; in WildGuard, the fine-tuned model shows the lowest drift of any fine-tuned condition (17.21) yet the worst behavioral degradation — low displacement does not prevent safety failure when the class-separating geometry erodes. Practically, this means CKA and Fisher score should be monitored throughout fine-tuning as primary safety health indicators, providing an early warning signal that behavioral output statistics cannot detect until collapse is complete.

Ambiguous Outputs as a Collapse Diagnostic. The 100% ambiguous rate for Granite Guardian's fine-tuned condition is not a neutral intermediate outcome but a signature of complete representational collapse: the destroyed classification boundary produces incoherent outputs that offer no safety protection and may pass silently through downstream safety checks that expect well-formed refusals or compliant responses. In multi-agent pipelines this is especially dangerous, as collapsed guard outputs become the inputs to downstream agents and can propagate unpredictably through the system. Safety evaluation protocols should treat ambiguous rate as a collapse diagnostic rather than a benign category, alongside standard utility benchmarks that provide no signal of geometric degradation.

Limitations. Our evaluation is conducted under a single-agent fine-tuning setting and does not model cascading failure in live multi-agent pipelines with real inter-agent communication. The safety probe covers 38 unique prompts across 5 harm categories; as evidenced by the Physical Harm anomaly in WildGuard (0% original and fine-tuned ASR, but 10% under FW-SSR), insufficient probe coverage of specific harm categories weakens subspace regularization in those directions, and category-stratified probe construction with broader coverage is needed for production deployment. The diagonal Fisher approximation ignores inter-direction correlations within the safety subspace, and the interaction between FW-SSR and deliberately adversarial fine-tuning data (Yang et al. 2023) requires separate investigation.

Conclusion

Benign fine-tuning catastrophically destroys safety geometry in purpose-built guard models– a severity exceeding prior findings on general-purpose LLMs– because concentrated safety representations are disproportionately targeted by task gradients. FW-SSR mitigates this through curvature-aware Fisher weighting and adaptive λ scheduling, recovering 75% refusal and CKA $= 0.983$ on Granite Guardian, and reducing WildGuard's ASR to 3.6% below the unmodified baseline. The central finding is that structural representational geometry– not pointwise activation magnitude– is the primary carrier of safety behavior, and CKA-based metrics should be first-class indicators in future guard model evaluation and agentic deployment pipelines.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation (Award No. 2451946) and the U.S. Nuclear Regulatory Commission (Award No. 31310025M0012). ChatGPT was utilized to assist with language editing and clarity improvements in this work. No content was generated related to technical results, data, code, or analysis.

Ethical Statement

This work studies a vulnerability in AI safety systems with the explicit goal of developing defenses. Harmful prompts in the safety probe dataset were authored by the researchers for evaluation purposes and are not publicly released. FW-SSR is a defensive technique; we do not anticipate dual-use concerns. No safety-compromised model weights are released.

References

Dettmers, T.; Pagnoni, A.; Fansi, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; and Fritz, M. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv preprint arXiv:2302.12173*.

Han, S.; Kim, K.; Youn, R.; Kim, J.; Longpre, S.; Haejun, L.; and Shin, J. 2024. WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In *Advances in Neural Information Processing Systems*.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S. F.; and Liu, L. 2024. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Advances in Neural Information Processing Systems*.

IBM Research. 2024. Granite Guardian. *arXiv preprint arXiv:2412.07724*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations. *arXiv preprint arXiv:2312.06674*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114: 3521–3526.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*, 3519–3529.

Min, R.; Qin, Z.; Zhang, N. L.; Shen, L.; and Cheng, M. 2024. Uncovering, Explaining, and Mitigating the Superficial Safety of Backdoor Defense. In *Advances in Neural Information Processing Systems*.

OpenSafetyLab. 2024. MD-Judge: A Multi-Dimensional Safety Judge for Open-Source Safety Evaluation of Large Language Models. *arXiv preprint arXiv:2406.17512*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint arXiv:2311.03658*.

Perez, F.; and Ribeiro, I. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. *arXiv preprint arXiv:2211.09527*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual Adversarial Examples Jailbreak Aligned Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv preprint arXiv:2310.03693*.

Shi, H.; Xu, Z.; Wang, H.; et al. 2024. Continual Learning of Large Language Models: A Comprehensive Survey. *ACM Computing Surveys*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-Following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6).

Wang, R.; et al. 2025. Simulated Ensemble Attack: Transferring Jailbreaks Across Fine-tuned Vision-Language Models. *arXiv preprint arXiv:2508.01741*.

Yang, X.; Wang, X.; Zhang, Q.; Petzold, L.; Wang, W. Y.; Zhao, X.; and Lin, D. 2023. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models. *arXiv preprint arXiv:2310.02949*.

Zhang, Z.; Sun, M.; Ye, X.; Wei, Y.; Peng, Y.; Chen, D.; Pang, H.; and Wu, F. 2024. How Alignment and Jailbreak Work: Explain LLM Safety Through Intermediate Hidden States. *arXiv preprint arXiv:2406.05644*.

Zhu, Y.; Liu, D.; Lin, Z.; Tong, W.; Zhong, S.; and Shao, J. 2025. The LLM Already Knows: Estimating LLM-Perceived Question Difficulty via Hidden Representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. In *arXiv preprint arXiv:2310.01405*.