BUILDING GEOMETRY-AWARE PROBES FOR FINE-GRAINED LLM SAFETY CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal activation probes offer an efficient approach to safety monitoring in large language models, but most existing work focuses on general harmfulness rather than fine-grained safety categories. We study how harmful categories are organized in representation space and how this geometry affects classification. Category directions relative to safe inputs exhibit a strong shared orientation while retaining category-specific variation. Categories that align more strongly with this shared geometry also tend to be harder to classify, with Spearman correlations of -0.691 for mean inter-category directional similarity and -0.682 for shared-axis alignment. Motivated by this observation, we introduce Spatial-Sign-Conditioned Angular Kernel (SSCAK) probing. SSCAK estimates shared within-class directional scatter using spatial signs, conditions the representation space accordingly, and learns nonlinear category boundaries on the unit hypersphere using an angular kernel. On WildGuard, across seven instruction-tuned models and three token-aggregation strategies, SSCAK consistently improves macro-F_1 over the corresponding standard probes. These results connect the geometry of internal safety representations to fine-grained classification and demonstrate the value of explicitly modeling within-class directional structure.

1 INTRODUCTION

Large language models increasingly rely on safety monitoring systems to identify harmful or policy-violating interactions. Dedicated guard models such as Llama Guard, WildGuard, AEGIS, and Qwen3Guard address this problem through external text classification (Inan et al., 2023; Han et al., 2024; Ghosh et al., 2024; Zhao et al., 2025a). In parallel, safety-relevant information can be read directly from internal activations. Activation probes provide a lightweight form of monitoring and have shown strong performance in harmful content detection, high-stakes interaction detection, out-of-distribution monitoring, and deployment-oriented safety systems (McKenzie et al., 2025; Jiao et al., 2026; Kramár et al., 2026; Cunningham et al., 2026). These advances establish internal representations as a useful source of safety signals.

Fine-grained safety classification presents a distinct challenge. Detecting that an input is harmful is often insufficient for moderation systems that must distinguish categories such as cyber misuse, privacy violations, fraud, discrimination, and physical harm. Recent studies suggest that these categories are not encoded as independent signals: harmfulness-related directions share substantial structure while retaining category-specific components (Zhao et al., 2025b; Shah et al., 2025; Park et al., 2026). This shared safety geometry is therefore directly relevant to the ability of activation probes to distinguish fine-grained categories.

We examine this relationship using the safe category as a common reference. Each harmful category is characterized by the direction of its mean representation relative to the safe mean. Among the 13 harmful categories in our analysis, 10 have more than half of their squared directional norm aligned with a shared reference axis, and the mean shared-axis fraction is 0.6566. Pairwise directional similarities nevertheless vary across categories, showing that a shared harmfulness orientation coexists with category-specific geometry.

We then connect this geometry to fine-grained classification. For each harmful category, we measure its average directional similarity to the remaining categories and its alignment with the shared reference axis. Both measures are negatively associated with per-category F_1: the Spearman correlations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are -0.691 and -0.682, respectively. Categories embedded more strongly in the shared geometry therefore tend to be harder for a conventional linear probe to distinguish. This relationship motivates a readout that explicitly accounts for directional structure.

To this end, we introduce Spatial-Sign-Conditioned Angular Kernel (SSCAK) probing. SSCAK decomposes deviations from each category center into magnitude and direction, then uses spatial signs to estimate shared within-class directional scatter. This scatter defines a common geometric scale for conditioning the representation space. The conditioned representations are normalized onto the unit hypersphere, where an angular RBF kernel captures nonlinear and potentially multimodal category structure. Pairwise kernel classifiers and probability coupling then produce the final multiclass prediction.

Our evaluation separates the proposed classifier from token aggregation. We apply the same classifier to mean, last-token, and attention pooling so that its contribution can be assessed independently of how token activations are combined. We evaluate models from the Gemma 2, Llama 3, Qwen3, and OLMo 3 families (Gemma Team et al., 2024; Grattafiori et al., 2024; Yang et al., 2025; Team OLMo et al., 2025) on WildGuard, AEGIS, Aegis2.0, OpenAI Moderation, and BeaverTails (Han et al., 2024; Ghosh et al., 2024; 2025; Markov et al., 2022; Ji et al., 2023). On WildGuard, SSCAK improves macro-F_1 over the corresponding standard probe in every reported model–aggregation comparison.

Our contributions are threefold:

1. **Geometry of fine-grained safety representations.** We show that harmful categories exhibit a substantial shared orientation relative to safe inputs while retaining heterogeneous category-specific directional structure.
2. **Geometry–performance connection.** We find that stronger inter-category directional similarity and greater shared-axis alignment are associated with lower category-level classification performance, directly linking shared safety geometry to probe behavior.
3. **Geometry-aware fine-grained probing.** We introduce SSCAK, which models shared within-class directional scatter with spatial-sign statistics and captures nonlinear category structure with an angular kernel. We evaluate the same classifier across multiple aggregation strategies, model families, and safety taxonomies.

2 Related Work

Internal Probes for Safety Monitoring. Dedicated guard models, including Llama Guard, WildGuard, AEGIS, and Qwen3Guard, assess safety from text and often expose explicit risk taxonomies (Inan et al., 2023; Han et al., 2024; Ghosh et al., 2024; Zhao et al., 2025a). A complementary line of work reads safety signals directly from internal activations. Activation probes have been studied for high-stakes interaction and harmful-content detection, deployment settings involving long contexts and distribution shifts, and production-oriented screening architectures (McKenzie et al., 2025; Jiao et al., 2026; Kramár et al., 2026; Cunningham et al., 2026). Nonlinear probes further capture higher-order feature interactions (Oldfield et al., 2026). Collectively, these studies show that internal activations can support lightweight monitoring across out-of-distribution inputs, long contexts, and multi-turn interactions. We extend this literature from general harmfulness detection to fine-grained categories and study how their representation geometry affects category-level readout.

Safety Representation Geometry. High-level concepts can form directions or low-dimensional structures in activation space (Zou et al., 2023; Park et al., 2025). Safety work first identified a refusal direction (Arditi et al., 2024) and later richer structures containing multiple directions and concept cones (Wollschläger et al., 2025). Recent studies distinguish harmfulness from refusal, find shared structure among harmfulness subconcepts, and decompose category representations into general-harm and category-specific components (Zhao et al., 2025b; Shah et al., 2025; Park et al., 2026). We connect this shared and category-specific geometry to fine-grained classification performance. Unlike work centered on refusal behavior or activation steering, our analysis uses a common safe reference to compare multiple risk categories and directly relates their directional organization to per-category readout difficulty. The resulting observations motivate the geometric conditioning used by SSCAK.

Fine-Grained Safety Taxonomies and Benchmarks. Fine-grained benchmarks provide complementary safety taxonomies: OpenAI Moderation covers real-world policy categories, WildGuard includes harmfulness, jailbreak, and refusal annotations, AEGIS and Aegis2.0 provide broad risk taxonomies, and BeaverTails supplies human-annotated question–answer pairs (Markov et al., 2022; Han et al., 2024; Ghosh et al., 2024; 2025; Ji et al., 2023). We use these resources to study how their categories are organized and read out from internal representations. Their different category definitions and levels of granularity also provide a useful test of whether the proposed readout transfers across safety taxonomies.

3 Geometry of Fine-Grained Safety Representations

Before developing readout methods for fine-grained safety classification, we first examine the geometric structure of different safety categories in the representation space of a frozen language model. Specifically, we focus on two related questions: whether the representation shifts of different harmful categories relative to the safe category contain a substantial shared directional component, and whether such geometric relationships are associated with the difficulty of fine-grained classification.

For a request x, we extract token representations $h_t(x)$ from a specified Transformer layer of the language model and obtain a request-level representation $z(x)$ through token-wise normalization and aggregation. Based on these representations, we study two related objects. The first is the **category direction**, defined by the displacement of each category mean relative to the mean of the safe category, which characterizes the global geometric relationships among safety categories in representation space. The second is the decision function learned by a supervised classifier on the same frozen representations, which allows us to examine whether these geometric relationships are associated with category-level recognition performance.

Section 3.1 begins with the geometry of category means. Using the safe category as a common reference, we analyze the components of harmful-category directions along a shared reference axis as well as their pairwise similarities. This analysis characterizes the extent to which a common orientation coexists with category-specific directional differences, and whether this structure is uniform across categories. Section 3.2 then trains a single-label multiclass probe on the same type of representations and re-estimates the category geometry using the probe's training samples. We use this setup to examine the relationship between directional similarity, shared-axis components, and per-category performance on held-out test data.

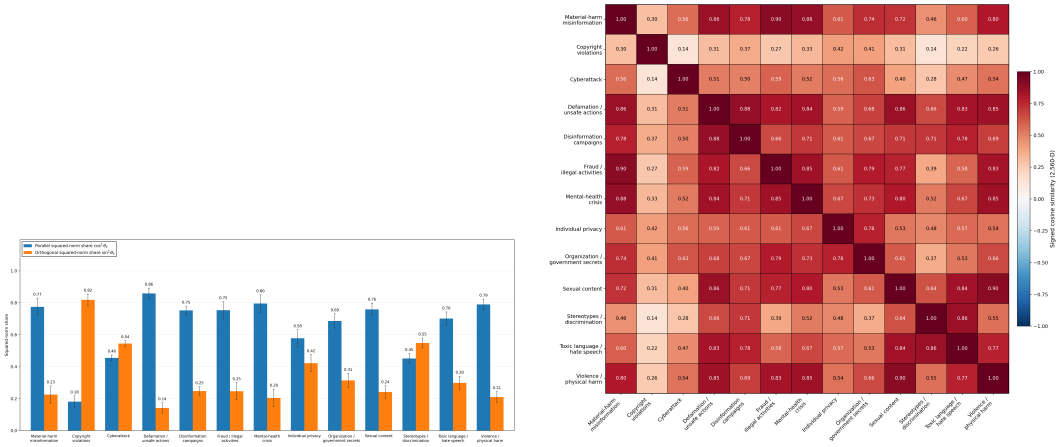
3.1 Shared Components of Category Directions

We first examine whether the representation shifts of different harmful categories relative to the safe category exhibit a common directional structure, and to what extent category-level differences are preserved within this shared geometry. We characterize category directions from two complementary perspectives: the fraction of their squared norm aligned with a shared reference axis, and their pairwise cosine similarities in the full representation space.

We conduct our analysis using `Qwen/Qwen3-4B-Instruct-2507` (Yang et al., 2025) on a class-balanced subset of the WildGuard training split (Han et al., 2024), covering 13 harmful categories and a safe category. For a request x, we extract residual-stream activations $h_t(x) \in \mathbb{R}^{2560}$ from a fixed Transformer layer. We observe that the first template token has a substantially larger representation norm than the remaining tokens, making the pooled representation sensitive to sequence length. To reduce the influence of large-norm tokens, we first ℓ_2-normalize each token activation along the hidden dimension and then average over all valid tokens in the complete chat template:

$$z(x) = \frac{1}{|T_x|} \sum_{t \in T_x} \frac{h_t(x)}{\max(\|h_t(x)\|_2, \varepsilon)}, \qquad \varepsilon = 10^{-12}. \tag{3.1}$$

Here, T_x denotes the valid tokens in the request together with the surrounding template tokens. This normalization reduces the contribution of a small number of high-norm tokens to the request-level mean representation, allowing $z(x)$ to more directly reflect the aggregate directional structure across tokens.



(a) Shared-axis and orthogonal squared-norm fractions by category.

(b) Pairwise signed cosine similarities between category directions.

Figure 1: (a) shows that 10 of the 13 harmful categories have a squared-norm fraction greater than 0.5 along the shared reference axis, with an average fraction of 0.6566 across categories. The signed projections $u^\top d_k$ are positive for all categories, indicating that their displacements from the safe mean are oriented toward the same side of the shared reference axis. (b) further reveals the overall relationships among category directions. All off-diagonal cosine similarities are positive. For example, the cosine similarity between Sexual content and Violence/physical harm is approximately 0.90.

Let μ_0 denote the mean representation of safe requests and μ_k the mean representation of the k-th harmful category. We define each category direction relative to the safe mean and construct a shared reference axis from the equally weighted harmful-category center:

$$\bar{\mu}_H = \frac{1}{H}\sum_{k=1}^{H}\mu_k, \qquad d_k = \frac{\mu_k - \mu_0}{\|\mu_k - \mu_0\|_2}, \qquad u = \frac{\bar{\mu}_H - \mu_0}{\|\bar{\mu}_H - \mu_0\|_2}. \tag{3.2}$$

Here, H is the number of harmful categories, d_k describes the orientation of category k relative to safe requests, and u captures their average orientation. We summarize this geometry using

$$p_k = (u^\top d_k)^2, \qquad C_{ij} = d_i^\top d_j. \tag{3.3}$$

Because u and d_k are unit vectors, p_k is the squared-norm share of d_k along the shared axis, while $1 - p_k$ is the share in its orthogonal complement. The signed projection $u^\top d_k$ indicates which side of the shared axis contains the category direction, and C_{ij} is the pairwise cosine similarity between categories.

Figure 1 shows that a shared orientation coexists with category-specific variation. Ten of the 13 harmful categories have $p_k > 0.5$, with a mean shared-axis fraction of 0.6566, and all signed projections $u^\top d_k$ are positive. Meanwhile, all off-diagonal similarities C_{ij} are positive but vary across category pairs. We next examine whether these geometric differences are associated with fine-grained classification performance.

3.2 Correlation Between Shared Geometry and Classification Performance

Section 3.1 reveals a geometric structure in which harmful categories share a common orientation while retaining category-specific directional variation. We next examine whether this geometry is related to fine-grained classification performance. We train a 14-class linear probe on the same frozen

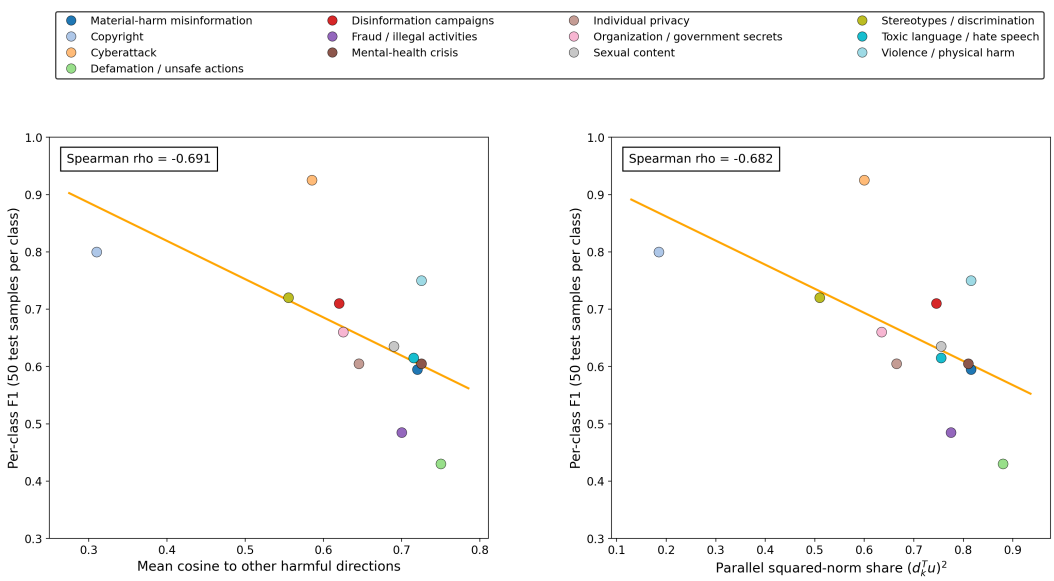


Figure 2: Relationship between shared category geometry and fine-grained classification performance. Each point represents one harmful category. Left: per-category F_1 versus mean inter-category cosine similarity S_k. Right: per-category F_1 versus shared-axis squared-norm fraction p_k. The annotations report Spearman rank correlations, and the orange lines show linear trends.

representations, estimate the category geometry using the probe's training samples, and evaluate per-category F_1 on held-out data.

Using the category directions from Section 3.1, we retain the shared-axis fraction p_k defined above and compute each category's mean cosine similarity to the remaining harmful categories:

$$S_k = \frac{1}{H-1} \sum_{\substack{j=1 \\ j \neq k}}^{H} d_k^\top d_j. \tag{3.4}$$

A larger S_k indicates greater directional similarity to other harmful categories, whereas a larger p_k indicates stronger alignment with the shared axis. Figure 2 shows that both measures are negatively associated with per-category F_1, with Spearman correlations of -0.691 for S_k and -0.682 for p_k.

The relationship also varies across categories. Copyright exhibits relatively low directional similarity and shared-axis alignment, with an F_1 of 0.800, whereas Defamation/unsafe actions has high values for both geometric measures and an F_1 of 0.432. Violence/physical harm, in contrast, combines relatively strong shared geometry with an F_1 of 0.747, highlighting heterogeneity across categories.

Overall, categories whose directions are more similar to those of other harmful categories and more strongly aligned with the shared reference axis tend to exhibit lower fine-grained classification F_1. Together with the geometric observations in Section 3.1, these results indicate an association between shared directional structure and fine-grained category recognition, motivating the category-specific readout developed in the next section.

4 Building a Category-Specific Classifier

The analysis in Section 3 shows that harmful categories exhibit a clear shared directional structure in representation space while retaining category-specific directional variation. Moreover, categories with stronger shared structure tend to have lower fine-grained classification performance. Motivated by this observation, we seek a readout that captures shared within-class structure while emphasizing directional differences across categories.

Our approach decomposes within-class variation into magnitude and direction, uses spatial-sign statistics to estimate shared directional scatter across categories, and conditions the representation space accordingly. After this shared geometric adjustment, we map the representations onto the unit hypersphere and use an angular kernel to learn nonlinear boundaries among fine-grained categories.

4.1 Problem Setup and Motivation

Given a single-label fine-grained safety classification dataset

$$\mathcal{D} = \{(q_i, y_i)\}_{i=1}^{N}, \tag{4.1}$$

where $y_i \in \{0, \dots, K-1\}$, we extract residual-stream representations from a fixed Transformer layer of the language model. For an input q, let $h_t(q)$ denote the representation of the t-th valid token. Following the representation construction in Section 3, we define $x(q) \in \mathbb{R}^d$ by first ℓ_2-normalizing each token representation and then averaging over valid tokens:

$$x(q) = \frac{1}{T_q} \sum_{t=1}^{T_q} \frac{h_t(q)}{\max\{\|h_t(q)\|_2, \varepsilon\}}, \qquad \varepsilon = 10^{-12}. \tag{4.2}$$

Here, T_q denotes the number of valid tokens for input q. Section 3 focuses on the geometry of category means, whereas classification boundaries also depend on within-class structure. To characterize this structure, we write the representation of category k as

$$X \mid Y = k = \mu_k + AV_k, \tag{4.3}$$

where μ_k is the category center, $A > 0$ denotes the magnitude of the displacement from the category center, and V_k is a unit direction satisfying

$$\|V_k\|_2 = 1. \tag{4.4}$$

This decomposition separates within-class variation into radial magnitude and direction. For fine-grained safety classification, we focus on differences in directional distributions across categories. Samples may vary in their displacement magnitude, while a single category may contain multiple discriminative directional modes.

Based on this representation, we use spatial-sign statistics to characterize within-class directional scatter. Each sample contributes according to its direction relative to the corresponding category center, providing a shared directional scale for subsequent classification.

4.2 Spatial-Sign-Conditioned Angular Kernel

Let $x_i = x(q_i)$ and $\mathcal{I}_k = \{i : y_i = k\}$. We first estimate the category means and the global mean. For each sample, we then compute its within-class residual and retain only its direction through the spatial-sign transformation (Visuri et al., 2000):

$$\begin{aligned} \mu_k &= \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} x_i, \quad & \mu &= \frac{1}{N} \sum_{i=1}^{N} x_i, \\ r_i &= x_i - \mu_{y_i}, & s_i &= \frac{r_i}{\max\{\|r_i\|_2, \varepsilon\}}, \\ S_w &= \frac{1}{N} \sum_{i=1}^{N} s_i s_i^\top. \end{aligned} \tag{4.5}$$

Thus, S_w summarizes shared directions of within-class variation while discarding their radial magnitudes. The constant ε is the same as in Equation equation 4.2 and prevents division by zero. We apply identity shrinkage and condition globally centered representations as

$$M = \rho I_d + (1-\rho) d S_w, \qquad z(x) = M^{-1/2}(x-\mu), \qquad 0 < \rho \leq 1. \tag{4.6}$$

For nondegenerate residuals, each spatial sign has unit norm and $\mathrm{tr}(S_w) = 1$, so dS_w has unit mean eigenvalue. The term ρI_d makes M positive definite and controls the strength of the conditioning. Consequently, $M^{-1/2}$ suppresses directions with large within-class directional variation relative to more stable directions.

After conditioning, we remove the remaining radial scale by mapping each nonzero representation onto the unit hypersphere. We then apply an RBF kernel whose distance is entirely determined by angular separation:

$$\begin{aligned} u(x) &= \frac{z(x)}{\max\{\|z(x)\|_2, \varepsilon\}}, \\ \kappa(x, x') &= \exp\left[-\gamma \|u(x) - u(x')\|_2^2\right] \\ &= \exp\left[-2\gamma\left(1 - u(x)^\top u(x')\right)\right]. \end{aligned} \tag{4.7}$$

Here, $\gamma > 0$ controls the decay of similarity with angular separation. For a fixed unit direction ν, $\kappa(u, \nu) = \exp(-2\gamma)\exp(2\gamma \nu^\top u)$, which has the same directional form as a von Mises–Fisher density with concentration 2γ (Mardia, 1975). A kernel classifier can combine support-vector-centered angular neighborhoods, enabling nonlinear and potentially multimodal decision regions on the hypersphere.

4.3 One-vs-One Classification and Pairwise Coupling

For each pair of categories $a < b$, we train a kernel SVM on $\mathcal{I}_{ab} = \mathcal{I}_a \cup \mathcal{I}_b$, assigning labels $t_i^{ab} \in \{-1, +1\}$. The pairwise decision functions and final prediction are

$$\begin{aligned} f_{ab}(x) &= \sum_{i \in \mathcal{I}_{ab}} \alpha_i^{ab} t_i^{ab} \kappa(x_i, x) + b_{ab}, \\ \widehat{y}(x) &= \arg \max_{k \in \{0, \ldots, K-1\}} p_k(x). \end{aligned} \tag{4.8}$$

We obtain $p(x)$ by calibrating the pairwise scores and combining the resulting probabilities through pairwise coupling (Wu et al., 2004). Thus, the complete readout conditions representations using shared within-class geometry and learns nonlinear category boundaries from their angular structure.

5 Experiments

5.1 Experimental Setup

Models and datasets. We evaluate our method on seven instruction-tuned language models: Gemma-2-2B-IT and Gemma-2-9B-IT (Gemma Team et al., 2024), Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-4B-Instruct-2507 and Qwen3-8B (Yang et al., 2025), and OLMo-3-7B-Instruct (Team OLMo et al., 2025). Experiments are conducted on five fine-grained safety datasets: WildGuard (Han et al., 2024), AEGIS (Ghosh et al., 2024), Aegis2.0 (Ghosh et al., 2025), OpenAI Moderation (Markov et al., 2022), and BeaverTails (Ji et al., 2023). All experiments follow the single-label classification setting defined in Section 4.1.

Token aggregation. We consider three strategies for aggregating token activations: mean pooling, last-token pooling, and attention pooling. Mean pooling averages token representations across the sequence, last-token pooling uses the representation at the final valid token position, and attention pooling learns a weighted combination of token representations (Ilse et al., 2018; McKenzie et al., 2025).

Classification. For each aggregation strategy, we compare the corresponding standard probe with the Spatial-Sign-Conditioned Angular Kernel classifier introduced in Section 4. The aggregation procedure is held fixed within each comparison, and only the downstream classification algorithm is varied. This produces paired comparisons for mean, last-token, and attention aggregation, allowing us to test whether the proposed classifier consistently improves fine-grained safety classification across representation constructions.

Evaluation. We use macro-F_1 as the primary evaluation metric. For each model–dataset pair, each standard-probe-versus-SSCAK comparison uses identical data splits and representations from the same model layer. Additional preprocessing, probe, and implementation details are provided in Appendix A.

5.2 Main Results

Table 1 reports macro-F_1 on WildGuard for the three token-aggregation strategies. Within each aggregation block, "Standard" denotes the corresponding baseline probe, while "+ SSCAK" applies our Spatial-Sign-Conditioned Angular Kernel classifier to the same aggregated representations.

Table 1: Macro-F_1 on WildGuard across models and token-aggregation strategies. SSCAK denotes the Spatial-Sign-Conditioned Angular Kernel classifier. The better result within each paired comparison is shown in bold.

	Mean pooling		Last-token pooling		Attention pooling	
Model	Standard	+ SSCAK	Standard	+ SSCAK	Standard	+ SSCAK
Gemma-2-2B-IT	0.7313	**0.8225**	0.6248	**0.6568**	0.7271	**0.8236**
Gemma-2-9B-IT	0.7317	**0.8371**	0.6619	**0.6889**	0.7528	**0.8359**
Llama-3.2-3B-Instruct	0.6979	**0.7824**	0.6529	**0.6970**	0.7168	**0.7933**
Llama-3.1-8B-Instruct	0.6988	**0.7616**	0.6217	**0.6803**	0.7460	**0.7980**
Qwen3-4B-Instruct-2507	0.7284	**0.8321**	0.6378	**0.6830**	0.7409	**0.8319**
Qwen3-8B	0.7365	**0.8396**	0.6272	**0.6780**	0.7384	**0.7996**
OLMo-3-7B-Instruct	0.7474	**0.8446**	0.7129	**0.7644**	0.7721	**0.8371**

Consistency across models and aggregation strategies. SSCAK improves all 21 comparisons in Table 1, with an average absolute gain of 0.0706 and a range of 0.0270–0.1054. Average gains are 0.0926 for mean pooling, 0.0750 for attention pooling, and 0.0442 for last-token pooling. The smaller last-token gain suggests that the classifier cannot fully recover information discarded by selecting only the final token. Nevertheless, every model improves under all three aggregation strategies, showing that the benefit is not confined to a particular architecture, scale, or representation construction.

Interaction with token aggregation. Attention is the strongest standard aggregation on average, but after SSCAK, mean and attention pooling both average 0.8171. Geometry-aware conditioning can therefore make simple mean aggregation competitive with learned aggregation. OLMo-3-7B achieves the strongest mean-pooling result (0.8446), while Qwen3-8B and Gemma-2-9B also exceed 0.83. This convergence between mean and attention pooling suggests that modeling directional class structure can be at least as important as learning a more expressive token-weighting function.

Table 2 compares SSCAK with dedicated guard models on the five safety datasets. The baselines include Llama Guard 3 (Grattafiori et al., 2024), Llama Guard 4 (Meta AI, 2025), Qwen3Guard (Zhao et al., 2025a), and Llama-3.1-Nemotron-Safety-Guard (Joshi et al., 2025). Our method uses mean-pooled activations from Qwen3-4B-Instruct-2507.

Table 2: Macro-F_1 comparison with dedicated guard models. Our SSCAK result uses mean-pooled activations from Qwen3-4B-Instruct-2507. The best result on each dataset is shown in bold.

Method	WildGuard	AEGIS	Aegis2.0	OpenAI Moderation	BeaverTails
Llama-Guard-3-8B	0.6718	0.6207	0.7109	0.7619	0.6375
Llama-Guard-4-12B	0.6923	0.6150	0.7006	0.7551	0.6380
Qwen3Guard-4B-Gen	0.7322	0.7521	0.7632	0.7239	**0.7507**
Qwen3Guard-8B-Gen	0.7602	0.7651	0.7519	0.7329	0.7221
Llama-3.1-Nemotron-Safety-Guard-8B-v3	0.7590	0.7516	0.7633	0.7513	0.7102
SSCAK (Qwen3-4B-Instruct-2507, mean)	**0.8321**	**0.8093**	**0.7921**	**0.7902**	0.6733

Comparison with dedicated guard models. SSCAK ranks first on four datasets, exceeding the strongest guard baseline by 0.0719 on WildGuard, 0.0442 on AEGIS, 0.0288 on Aegis2.0, and 0.0283 on OpenAI Moderation. On BeaverTails, Qwen3Guard-4B-Gen instead leads by 0.0774. This exception highlights the influence of dataset-specific taxonomies and motivates the cross-taxonomy directions discussed in Section 6. More broadly, the four wins indicate that internal activation readouts can be competitive with dedicated text-based guard models despite using a frozen base model and a lightweight downstream classifier.

6 Conclusion and Limitations

Conclusion. Harmful categories share a common orientation while retaining category-specific variation, and stronger shared geometry is associated with lower category-level performance. Motivated by this finding, SSCAK conditions shared within-class directional scatter and learns nonlinear angular boundaries, consistently improving the corresponding standard probes across seven models and three aggregation strategies on WildGuard.

Limitations. Safety taxonomies differ in definition and granularity and may not match a model's internal organization: samples within a category can be heterogeneous, while different categories may share structure. Consequently, performance may depend on how labels are filtered, merged, and balanced. Our formulation is also restricted to single-label classification and does not directly model inputs associated with multiple simultaneous risks.

Future Work. Future work can align taxonomies across datasets and use representation geometry to discover, refine, or reorganize latent safety categories. Extending SSCAK to multi-label and hierarchical settings, and evaluating whether the learned geometry transfers across datasets and model families, are also promising directions.

AI Use Statement

We used generative AI tools for drafting, language polishing, LaTeX formatting, and literature search, but not for data generation or experiment execution. Two authors reviewed all AI-assisted content, and the authors take responsibility for the final manuscript.

Ethics Statement

This study uses public models and datasets and involves no new human-subject or private data. Safety datasets may expose researchers to disturbing content and should be handled carefully. Although intended for defensive monitoring, the method may have dual-use implications. We report only aggregate results and no harmful examples.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Andy Dau, Alek Dimitriev, Logan Howard, Yijin Hua, Rob Gilson, Mu Lin, Christopher Liu, Vladimir Mikulik, Rohit Mittapalli, Clare O'Hara, Jin Pan, Nikhil Saxena, Alex Silverstein, Yue Song, Giulio Zhou, Jan Leike, Jared Kaplan, Ethan Perez, and Mrinank Sharma. Constitutional Classifiers++: Efficient production-grade defenses against universal jailbreaks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eNvsH5Ye2V.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. AEGIS: Online adaptive AI content safety moderation with ensemble of LLM experts. *arXiv preprint arXiv:2404.05993*, 2024. doi: 10.48550/arXiv.2404.05993. URL https://arxiv.org/abs/2404.05993.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. *arXiv preprint arXiv:2501.09004*, 2025. doi: 10.48550/arXiv.2501.09004. URL https://arxiv.org/abs/2501.09004.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.48550/arXiv.2406.18495. URL https://arxiv.org/abs/2406.18495.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human–AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. doi: 10.48550/arXiv.2312.06674. URL https://arxiv.org/abs/2312.06674.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1072. URL https://arxiv.org/abs/2307.04657.

Difan Jiao, Yilun Liu, Ye Yuan, Zhenwei Tang, Linfeng Du, Haolun Wu, and Ashton Anderson. LLM safety from within: Detecting harmful content with internal representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39711–39727, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1844. URL https://aclanthology.org/2026.acl-long.1844/.

Raviraj Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Long, Sanjay Singh Chauhan, and Niranjan Wartikar. CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications. *arXiv preprint arXiv:2508.01710*, 2025. doi: 10.48550/arXiv.2508.01710. URL https://arxiv.org/abs/2508.01710.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for Gemini. *arXiv preprint arXiv:2601.11516*, 2026. doi: 10.48550/arXiv.2601.11516. URL https://arxiv.org/abs/2601.11516.

Kanti V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(3):349–371, 1975. doi: 10.1111/j.2517-6161.1975.tb01550.x.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274*, 2022. doi: 10.48550/arXiv.2208.03274. URL https://arxiv.org/abs/2208.03274.

Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. Detecting high-stakes interactions with activation probes. In *Advances in Neural Information Processing Systems*, volume 38, pp. 141453–141491, 2025. doi: 10.52202/085713-4249.

Meta AI. Llama Guard 4 model card. Hugging Face model card, 2025. URL https://huggingface.co/meta-llama/Llama-Guard-4-12B.

James Oldfield, Philip Torr, Ioannis Patras, Adel Bibi, and Fazl Barez. Beyond linear probes: Dynamic safety monitoring for language models. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2509.26238. URL https://arxiv.org/abs/2509.26238.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2406.01506. URL https://arxiv.org/abs/2406.01506.

Soyeon Park, Seogyeong Jeong, Sunwoo Kim, and Alice Oh. The role of fine-grained harm signals in LLM safety. *arXiv preprint arXiv:2609.19366*, 2026. doi: 10.48550/arXiv.2609.19366. URL https://arxiv.org/abs/2609.19366.

McNair Shah, Saleena Angeline, Adhitya Rajendra Kumar, Naitik Chheda, Kevin Zhu, Vasu Sharma, Sean O'Brien, and Will Cai. The geometry of harmfulness in LLMs through subconcept probing. *arXiv preprint arXiv:2507.21141*, 2025. doi: 10.48550/arXiv.2507.21141. URL https://arxiv.org/abs/2507.21141.

Team OLMo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, et al. OLMo 3. *arXiv preprint arXiv:2512.13961*, 2025. doi: 10.48550/arXiv.2512.13961. URL https://arxiv.org/abs/2512.13961.

Samuli Visuri, Visa Koivunen, and Hannu Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575, 2000. doi: 10.1016/S0378-3758(00)00199-3.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025. doi: 10.48550/arXiv.2502.17420. URL https://arxiv.org/abs/2502.17420.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004. URL https://jmlr.org/papers/v5/wu04a.html.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng,

Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025a. doi: 10.48550/arXiv.2510.14276. URL https://arxiv.org/abs/2510.14276.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140283–140318, 2025b. doi: 10.52202/085713-4688.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405. URL https://arxiv.org/abs/2310.01405.

A Experimental Details

A.1 Dataset Preparation

We preprocess each dataset to obtain a class-balanced, single-label classification task. The preprocessing pipeline includes deduplication, removal of samples with inconsistent or unusable labels, and the merging of labels where necessary to obtain a coherent category set. Splits are created after filtering and are balanced at the category level.

For WildGuard, the resulting taxonomy contains 14 categories, including one safe category. We sample 300 training examples, 50 validation examples, and 50 test examples per category. During manual and rule-based filtering, we remove examples whose content is inconsistent with the assigned category label.

For both AEGIS and Aegis2.0, we use 100 training examples, 30 validation examples, and 30 test examples per retained category. Categories with too few usable examples are excluded; these include rare categories concerning threats and sexual content involving minors. Aegis2.0 originally permits multiple labels per example, so we retain only examples associated with exactly one safety category. After preprocessing, Aegis2.0 contains 11 categories, including one safe category.

OpenAI Moderation is used only as an external test set, with 50 examples per category, and is not used for training or hyperparameter selection. For BeaverTails, we use 300 training examples, 50 validation examples, and 50 test examples per category.

A.2 Activation Extraction Layers

All language-model parameters remain frozen. We extract residual-stream activations from one fixed layer of each model. Layer indices are zero-based and are listed in Table 3.

Table 3: Residual-stream layers used for activation extraction. Layer indices are zero-based.

Model	Layer index
Gemma-2-2B-IT	13
Gemma-2-9B-IT	19
Llama-3.2-3B-Instruct	16
Llama-3.1-8B-Instruct	14
Qwen3-4B-Instruct-2507	20
Qwen3-8B	17
OLMo-3-7B-Instruct	18

A.3 Standard Probe Architectures

Following the standard activation-probe definitions of McKenzie et al. (2025), we adapt the binary probe formulations to our multiclass setting by instantiating one logit per category. We represent the residual activations of an input as

$$A = [a_1, \ldots, a_S]^\top \in \mathbb{R}^{S \times D}, \tag{A.1}$$

where S is the number of valid tokens and D is the hidden dimension. For category c, the mean and last-token probe logits are

$$\begin{aligned} f_c^{\text{mean}}(A) &= \theta_c^\top \left(\frac{1}{S} \sum_{s=1}^{S} a_s \right) + b_c, \\ f_c^{\text{last}}(A) &= \theta_c^\top a_S + b_c. \end{aligned} \tag{A.2}$$

Thus, the mean probe first averages the token representations and applies a linear classification layer, whereas the last-token probe applies the linear layer only to the final valid token representation. The attention probe uses a category-specific query direction to weight token-level scores:

$$\begin{aligned} \alpha_{s,c} &= \frac{\exp(\theta_{q,c}^\top a_s)}{\sum_{j=1}^{S} \exp(\theta_{q,c}^\top a_j)}, \\ f_c^{\text{attn}}(A) &= \sum_{s=1}^{S} \alpha_{s,c}\, \theta_{v,c}^\top a_s + b_c. \end{aligned} \tag{A.3}$$

Here, $\theta_{v,c}$ directly serves as the category-classification direction; no additional classification head is applied after attention aggregation. For all three probes, the category probability is $p_c(A) = \sigma(f_c(A))$. We train the probes using binary cross-entropy on one-hot category targets and predict the category with the largest output probability.

A.4 Optimization and Compute Environment

The standard probes are optimized with AdamW. The SVM penalty C, angular kernel parameter γ, and regularization parameters of SSCAK are selected through grid search using the available validation partitions. OpenAI Moderation remains strictly held out during this process. Experiments are run on an NVIDIA GeForce RTX 5090 GPU using Python 3.12.3 and PyTorch 2.11.0+cu130.