

# The Geometry of Alignment Collapse: When Fine-Tuning Breaks Safety

Max Springer<sup>1,\*</sup>, Chung Peng Lee<sup>1</sup>, Blossom Metevier<sup>1</sup>, Jane Castleman<sup>1</sup>,  
Bohdan Turbal<sup>1</sup>, Hayoung Jung<sup>1</sup>, Zeyu Shen<sup>1</sup>, Aleksandra Korolova<sup>1</sup>

<sup>1</sup>Princeton University

\*Corresponding Author: maxspringer@princeton.edu

Fine-tuning aligned language models on entirely benign tasks (e.g. math tutoring) unpredictably degrades safety guardrails, even when training data contains no harmful content and developers have no adversarial intent. We show that the prevailing explanation, that fine-tuning updates should be orthogonal to safety-critical directions in high-dimensional parameter space, offers false reassurance: we show that this orthogonality is structurally unstable and collapses under the very dynamics of gradient descent. We then resolve this through a novel geometric analysis, proving that alignment concentrates in low-dimensional subspaces with sharp curvature, creating a brittle structure that first-order methods cannot detect or defend. While initial fine-tuning updates may indeed avoid these subspaces, the curvature of the fine-tuning loss generates second-order acceleration that systematically steers trajectories into alignment-sensitive regions—an effect invisible to all existing defenses. We formalize this mechanism through the **Alignment Instability Condition (AIC)**, three geometric properties that, when jointly satisfied, lead to safety degradation. Our main result establishes a quartic scaling law: alignment loss grows with the fourth power of training time, governed by the sharpness of alignment geometry and the strength of curvature coupling between the fine-tuning task and safety-critical parameters. These results expose a structural blind spot in the current safety paradigm. The dominant approaches to safe fine-tuning of null-space projections, gradient filtering, and first-order constraints address only the initial snapshot of a fundamentally dynamic problem. Alignment fragility is not a bug to be patched; it is an intrinsic geometric property of gradient descent on curved manifolds. Our results motivate the development of curvature-aware methods, and we hope will further enable a shift in alignment safety analysis from reactive red-teaming to predictive diagnostics for open-weight model deployment.

Last updated: February 18, 2026



## 1 Introduction

Fine-tuning a language model on a dataset of math problems breaks its safety guardrails [12]. Training on creative writing examples erodes medical advice safeguards [1, 64]. Adapting for code generation compromises refusal mechanisms [6]. These are not hypothetical risks—they are documented, reproducible phenomena that occur even when fine-tuning datasets contain no harmful content and developers have no adversarial intent [6, 24, 25, 44]. The central puzzle of alignment preservation is thus:

*Why does adapting a model for one capability seemingly unpredictably degrade unrelated safety properties?*

The question is urgent because fine-tuning has become the dominant paradigm for deploying language models. In domains where technical expertise, compute resources, and specialized data are limited, fine-tuning open weight models represents the primary path to leverage AI capabilities for practitioners. The ecosystem of open-weight models has expanded rapidly [7, 9, 10, 31, 40], enabling innovation but also creating vulnerability. Current defenses are largely reactive, targeting specific attack patterns [18] or imposing first-order gradient constraints [16, 39, 65]

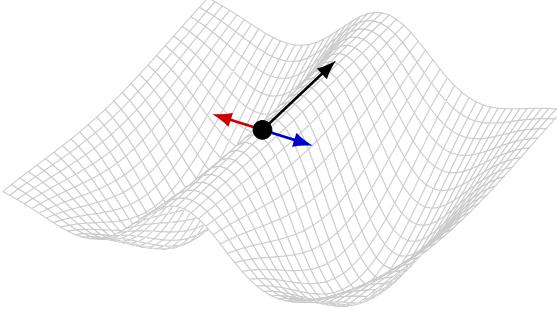


Figure 1: Local alignment instability under fine-tuning dynamics. An illustrative loss landscape showing a fine-tuning trajectory (black arrow) evolving near a ridge separating basins of low alignment utility. Although the initial gradient direction is nearly tangent to the ridge, induced acceleration toward a high-curvature alignment-sensitive direction (red/blue arrow), leads to departure.

that easily break even without adversarial intent [26, 63]. Without understanding the underlying mechanisms of alignment degradation, we cannot build robust safeguards.

**The Paradox of High-Dimensional Orthogonality.** Empirical evidence reveals a striking pattern: alignment appears to reside in a small subspace of the parameter space. [60] demonstrates that pruning just 3% of specific weights triggers catastrophic safety collapse, while removing similar proportions of weights from other parts of the model leaves its capabilities intact. We call these directions *alignment-sensitive*: parameter updates along them degrade safety, while updates in orthogonal directions should preserve it. This suggests alignment is governed by a low-dimensional set of sensitive directions. In a model with billions of parameters, standard geometric intuition predicts that fine-tuning updates for unrelated tasks should be nearly orthogonal to these directions, thus, safe [5, 35]. Yet degradation occurs frequently across diverse tasks and model families, exhibiting consistent patterns [6, 25, 44].

Existing theoretical frameworks do not fully resolve this paradox. Approaches based on data overlap [25, 26, 63] cannot fully explain why superficially similar tasks have vastly different safety impacts. First-order geometric constraints assume static sensitivity subspaces and fail to account for how fine-tuning trajectories evolve over time [5, 45]. Null-space methods [39, 65] presuppose we can identify and avoid alignment-sensitive directions but offer no explanation for why avoidance fails in practice.

**Our Contributions.** We resolve this paradox by developing a theory of alignment degradation grounded in the curvature of trajectories in parameter space. Our key insight is that while initial fine-tuning updates may indeed be orthogonal to alignment-sensitive directions, this orthogonality is unstable under the dynamics of gradient descent. The curvature of the fine-tuning objective—captured by second-order terms in the loss landscape—systematically bends trajectories into alignment-sensitive subspaces, even when first-order updates point away from them.

We formalize this mechanism through the **Alignment Instability Condition (AIC)**, which identifies three interacting structural properties arising from the model’s geometry and the fine-tuning task.

1. **Low-Rank Sensitivity:** alignment for a given skill (or capability) concentrates in a small subspace with sharp curvature, characterized by the leading eigenvectors of the Fisher information matrix.
2. **Initial Orthogonality:** fine-tuning gradients for unrelated tasks have negligible first-order projection onto this subspace, creating the illusion of safety.
3. **Curvature Coupling:** second-order terms in the gradient flow induce nontrivial acceleration into the alignment-sensitive subspace, forcing trajectories to drift despite initial orthogonality.

Under the AIC, we prove that alignment degradation is not merely possible, but structurally inevitable within the trust region of the aligned model. Our main theoretical results are:

**Theorem 1.1** (Informal Version of Theorem 6.1). *Movement in the alignment-sensitive subspace  $M_i \subseteq \mathbb{R}^n$  incurs quadratic utility loss,  $\Omega(\lambda \|P_i(\Delta\theta)\|^2)$  where  $\lambda$  is the minimum curvature in  $M_i$  and  $P_i$  projects onto this subspace.*

**Theorem 1.2** (Informal Version of Theorem 6.2). *Along gradient flow trajectories,  $\theta(t)$  where  $\theta(0) = \theta^*$ , satisfying the AIC, projection on  $M_i$  grows quadratically:  $\|P_i(\theta(t) - \theta^*)\| = \Omega(\gamma t^2)$ , where  $\gamma$  quantifies curvature coupling.*

The two main theorems combine to yield a quartic scaling law for alignment degradation within our utility-based framework (defined in Section 3).

**Corollary 1.3** (Informal Version of Corollary 6.3). *Combining geometric loss with quadratic drift yields alignment degradation scaling as  $\Omega(\lambda\gamma^2 t^4)$  in early training.*

The quartic scaling law provides a principled explanation for rapid, seemingly sudden safety failures observed in practice [20, 42, 43, 60]. Even small amounts of curvature coupling ( $\gamma$ ) compound dramatically through the fourth power, causing trajectories to escape safety basins faster than first-order analysis would predict.

**Implications.** Our theory reveals that alignment fragility is not an artifact of adversarial attacks or fixable through better data curation. Rather, it is an intrinsic geometric property of fine-tuning in high-dimensions. The AIC explains why: (1) benign tasks break safety, (2) low-rank adaptations fail, and (3) first-order defenses are insufficient. Our results necessitate a fundamental rethinking of alignment preservation strategies. In Section 7, we empirically validate these predictions, showing that our notion of geometric overlap provides a principled explanation for degradation of alignment.

## 2 Related Work

Our work is positioned at the intersection of several research domains: the geometric analysis of neural network parameter spaces, the empirical study of alignment fragility, and the theoretical frameworks of multi-task and meta-learning. We here discuss related studies which were not covered in the main text.

**Geometric and Manifold-Based Approaches.** The conceptualization of neural network representations and parameter spaces through a geometric lens is well-established [2, 47, 54]. This line of inquiry is broadly motivated by the manifold hypothesis [46], which posits that high-dimensional data, such as inputs for unsupervised learning or distinct classes for classification, concentrates near low-dimensional manifolds separated by regions of low density. More recently, this geometric perspective has been directly applied to the process of fine-tuning [38]. Specifically, Low-Rank Adaptation (LoRA) has been theoretically framed as a problem of manifold geometry [29]. This aligns with foundational work in numerical optimization that analyzes fixed-rank matrices as an embedded geometry, even outside the context of model fine-tuning [46].

While these approaches provide valuable intuition, they often treat alignment geometry as static. Recent work has attempted to identify null-spaces to the alignment weight subspace in an effort to constrain fine-tuning [65]. However, these methods typically assume a flat local geometry that does not account for the empirical fragility of alignment in high-dimensions—a gap our work addresses by integrating this geometric view with a probabilistic analysis of curvature-induced drift.

**Alignment Drift and Critical Parameters.** Empirical research confirms that fine-tuning modifies only a small subset of the representational subspace, presumably preserving the pre-trained model’s broad capabilities. However, this preservation is asymmetric: while general capabilities may persist, aligned guardrails are uniquely brittle. [26] demonstrate that this degradation is systematic, showing that the impact of fine-tuning varies dramatically depending on the specific task structure, a phenomenon further categorized in broader surveys of trustworthy AI [36]. Critically, this degradation occurs even on seemingly benign tasks such as solving math word problems [12] or standard instruction following [24, 25, 44] where seemingly no adversarial intent exists.

More critically, [16] introduce the “DriftCheck” dataset, demonstrating that even minor LoRA updates induce significant alignment drift. In fine-tuning with this dataset, the identify alignment critical parameters that are most relevant to maintaining safety behavior, finding that an overlap between the fine-tuning data and safety distributions accelerates this drift by overwriting fragile weight regions. This observation connects directly with the safety relevant weights hypothesis of [60], which assesses the brittleness of safety alignment via pruning, showing that removing specific weights triggers rapid safety collapse. [25] and [24] further demonstrate that fine-tuning on benign data can compromise safety even without adversarial intent. Our proposal seeks to explore the theoretical mechanism, which

we refer to as the **Alignment Instability Condition**, that explains why these critical parameters are susceptible to drift even when the ecosystem of open-model development is expanding rapidly [40].

**Theoretical Frameworks for Skill Emergence and Adaptation.** Our decomposition of alignment into distinct utility functions is grounded in recent theoretical advances regarding skill emergence in large language models. [4] formally model this phenomenon, proposing that complex skills emerge from the scaling of random feature combinations that allow the model to distinguish between increasingly granular tasks. This skill-based view is further supported by [37], who provide a kernel-based analysis of how fine-tuning modifies these underlying feature mappings. Our work adopts this foundational perspective by viewing alignment features not as global constraints but as distinct, emergent skills. However, we reframe the problem from acquisition to preservation. While [4] explain how these high-dimensional skills arise, our geometric theory addresses the inverse problem: why specific skills possess a unique structural fragility that makes them susceptible to catastrophic forgetting even when general capabilities remain intact. This fragility is further understood by examining the intersection of our work with meta-learning theory. The standard paradigm for multi-task learning, as established in [19, 56], posits that diverse tasks can be solved by learning a single, shared low-dimensional representation, upon which task-specific linear heads are trained. In this idealized setting, fine-tuning is safe because the underlying representation is fixed. However, practical fine-tuning violates this assumption by updating the representation itself. [11] address this by relaxing the requirement to an approximate shared representation, where task-specific parameters are constrained to remain within a small distance of the central meta-parameters. While [11] prove that this relaxation allows for effective adaptation, our AIC reveals a critical blind spot in this framework regarding safety. We demonstrate that on curved manifolds, the allowable deviation cannot be isotropic. Due to the high curvature associated with alignment skills, even valid updates that satisfy the distance constraint for a new task can induce large, non-linear drifts in the safety subspace. Thus, our work challenges the sufficiency of the approximate shared representation hypothesis for safety-critical systems, showing that geometric curvature couples benign task updates with more harmful alignment degradation in ways that standard meta-learning bounds do not predict.

### 3 Model & Preliminaries

We consider a language model parametrized by  $\theta \in \Theta \subseteq \mathbb{R}^n$ , defining a conditional distribution, or policy,  $\pi_\theta(y|x)$  over responses  $y \in \mathcal{Y}$  given prompts  $x \in \mathcal{X}$ . Unless otherwise stated, we let  $\|\cdot\|$  denote the Euclidean norm of a vector and  $\|\cdot\|_{\text{op}}$  denote the operator norm.

#### 3.1 Alignment as Skill Utility

A key obstacle to theoretical analysis is that alignment is not monolithic. A model simultaneously maintains multiple capabilities that can degrade independently during fine-tuning. We therefore decompose alignment into distinct *skills* (or capabilities), following recent work on skill emergence in language models [4, 37].

**Definition 3.1** (Alignment Skill and Utility). *Let  $\mathcal{S} = \{S_1, \dots, S_m\}$  be a set of alignment skills. Each skill  $S_i$  is associated with a reference distribution  $\mathcal{D}_i$  over interaction pairs  $(x, y)$ . We define the **utility** of the model  $\theta$  for skill  $S_i$ ,*

$$u_i(\theta) := \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\log \pi_\theta(y|x)].$$

For a fixed base model,  $\theta^*$ , we define the utility loss (or alignment degradation) for skill  $S_i$  under parameter update  $\Delta\theta$  as  $\Delta u_i(\theta) := u_i(\theta^*) - u_i(\theta)$ . This definition grounds model alignment in probabilistic terms: a model is well-aligned for skill  $S_i$  if it assigns high likelihood to responses deemed ideal under  $\mathcal{D}_i$ .

#### 3.2 Local Geometry and Fisher Information

To analyze alignment preservation geometrically, we adopt an idealized but analytically convenient assumption that the base model is well-optimized for the alignment skills under consideration.

**Assumption 1** (Skill Optimality). *The base model  $\theta^*$  is optimal for every alignment skill  $S_i$ :  $\pi_{\theta^*}(y|x) = \mathcal{D}_i(y|x)$  for all  $x \in \text{supp}(\mathcal{D}_i)$  and all  $y$  with  $\mathcal{D}_i(y|x) > 0$ .*

This assumption states that prior to fine-tuning, the base model exactly satisfies the developer-specified alignment objectives. While idealized, similar well-defined assumptions are standard in theoretical analyses to obtain clean local characterizations of learning dynamics. In our setting, Assumption 1 ensures that  $\theta^*$  is stationary for each skill’s utility and that local alignment degradation admits a transparent reduction to a KL divergence [14]. We emphasize that Assumption 1 is stronger than necessary and is adopted for clarity of presentation. Our main results extend under weaker local conditions (see Appendix C).

**Lemma 3.2** (Alignment Loss as KL Divergence). *Given Assumption 1, after fine-tuning to obtain model  $\theta$ , the degradation in utility for any skill  $S_i$  is equal to the expected KL divergence between the policies, measured over the distribution  $\mathcal{D}_i$ :*

$$\Delta u_i = \mathbb{E}_{x \sim \mathcal{D}_i(x)} [\text{D}_{\text{KL}}(\pi_{\theta^*}(y|x) || \pi_\theta(y|x))].$$

The proof follows from the definition of utility and is deferred to Appendix A. This result highlights that alignment stability is determined by the model’s policy divergence on the *alignment* distribution  $\mathcal{D}_i$ , which may differ significantly from the fine-tuning distribution. In practice, alignment comprises multiple skills (safety refusal, medical accuracy, factual correctness, etc.). Our analysis focuses on individual skill degradation, where overall alignment fails when any critical skill degrades below acceptable thresholds.

**Trust Regions and Local Quadratic Geometry.** A common safeguard, such as those in PPO [49] or TRPO [48], constrain updates within a KL trust region around the base policy:

$$D_{\text{KL}}(\pi_{\theta^*} || \pi_\theta) \leq \Delta,$$

where  $\Delta > 0$  is a small positive constant. While such constraints were initially designed for optimization stability, in alignment settings they are often interpreted as preventing harmful deviation from a safe reference model [62]. Geometrically, this restricts fine-tuning to a neighborhood where the parameter space is locally quadratic. We quantify this curvature using the Fisher Information Matrix (FIM). Let  $\mathbf{F}_x(\theta^*) = \mathbb{E}_{y \sim \pi_{\theta^*}(\cdot|x)} [\nabla_\theta \log \pi_{\theta^*}(y|x) \nabla_\theta \log \pi_{\theta^*}(y|x)^\top]$  be the Fisher matrix for input  $x$ . We further define the FIM specific to skill  $S_i$  as  $\mathbf{F}_i(\theta) := \mathbb{E}_{x \sim \mathcal{D}_i} [\mathbf{F}_x(\theta)]$ . Assuming the model is  $C^3$  smooth, the second-order Taylor expansion of the KL divergence yields the following local form for alignment loss:

**Proposition 3.3** (Local Geometric Form). *For sufficiently small displacements  $\Delta\theta = \theta - \theta^*$ ,*

$$\Delta u_i = \frac{1}{2} \Delta\theta^\top \mathbf{F}_i(\theta^*) \Delta\theta + O(\|\Delta\theta^3\|). \quad (1)$$

Thus, in the small-step regime enforced by KL trust regions, skill degradation is locally dominated by the projection of the update  $\Delta\theta$  onto the high-curvature directions of  $\mathbf{F}_i(\theta^*)$ .

### 3.3 Low-Rank Sensitivity Structure

The key geometric insight is that not all directions in parameter space affect alignment equally. Empirical studies of neural network loss landscapes reveal that Hessians (and FIMs) typically have spiked spectra: a few large eigenvalues dominate, while most directions have near-zero curvature [47]. Recent empirical work also shows that model weights share a low-rank joint subspace across different tasks [32]. This further motivates a low-rank decomposition of alignment sensitivity.

**Definition 3.4** (Alignment Sensitivity Subspaces). *The alignment sensitivity subspace  $M_i$  for skill  $S_i$  is the span of the  $d$ -leading eigenvectors of  $\mathbf{F}_i(\theta^*)$ .*

Stemming from this definition, let  $P_i : \mathbb{R}^n \rightarrow M_i$  denote the Euclidean orthogonal projection onto  $M_i$ , and  $P_i^\top = I - P_i$ , the projection onto its orthogonal complement. Within our framework, directions in  $M_i$  are critical for alignment, with small movements incurring large utility loss due to high curvature, whereas directions in the null space are locally benign. This decomposition formalizes the empirical observations that alignment resides in small subspaces [5, 35, 60].

## 4 Why Flat Geometry Fails

Section 3 established a local geometric picture: utility loss is controlled by projections onto low-dimensional sensitivity subspaces, and KL trust regions restrict fine-tuning to neighborhoods where this quadratic approximation holds. This suggests a powerful alignment-preservation guarantee: in high dimensions, unrelated fine-tuning updates should be nearly orthogonal to any fixed low-rank subspace, making degradation exponentially unlikely. We formalize this intuition as a probabilistic **null model**, then contrast it against empirical evidence showing that real fine-tuning trajectories violate these predictions systematically.

### 4.1 Benign Fine-Tuning in the Flat Regime

To model benign fine-tuning without additional structure, we idealize the update direction as a random unit vector in parameter space.

**Theorem 4.1** (Benign Fine-Tuning under Flat Geometry). *Fix an alignment skill  $S_i$  and assume there exists  $r > 0$  such that for all  $\|\Delta\theta\| \leq r$ , we have that  $\Delta u_i = \frac{1}{2} \Delta\theta^\top \mathbf{F}_i(\theta^*) \Delta\theta + O(\|\Delta\theta\|^3)$ . Let  $g \sim \text{Unif}(\mathbb{S}^{n-1})$  be a random unit vector, and consider the update  $\Delta\theta = \eta g$  where  $\eta > 0$  is sufficiently small. Then: (1) the expected utility loss satisfies*

$$\mathbb{E}_{g \sim \text{Unif}(\mathbb{S}^{n-1})} [\Delta u_i] = \frac{\eta^2}{2n} \text{Tr}(\mathbf{F}_i(\theta^*)) + O(\eta^3),$$

(2) for any rank- $d$  orthogonal projector  $P$ , we have  $\|P(g)\|^2 = \Theta(d/n)$ , with probability at least  $1 - 2e^{-c\varepsilon^2 d}$ . This further implies that

$$\Delta u_i = O\left(\frac{d\eta^2}{n}(1+\varepsilon)\lambda_{\max}(\mathbf{F}_i(\theta^*))\right).$$

The proof uses rotation invariance of the uniform distribution on the sphere and concentration of  $\chi^2$  random variables (Appendix A).

This theorem formalizes the high-dimensional safety intuition: when the utility loss is well-approximated by a fixed quadratic form and fine-tuning updates are unstructured, the expected degradation scales with the average curvature  $\text{Tr}(\mathbf{F}_i(\theta^*))/n$  and the probability of significant overlap with any fixed low-dimensional sensitivity subspace is exponentially small in its dimension. In this regime, alignment degradation is suppressed both in expectation and with high probability by the ambient dimensionality of the parameter space. Consequently, if harmful or alignment-critical directions are confined to a small subspace and fine-tuning updates are effectively random, narrow fine-tuning should be benign. However, this conclusion rests on two idealizations: (1) updates are unstructured (modeled as random), and (2) the Fisher geometry is static throughout training. As we show next, both assumptions fail in practice.

### 4.2 Empirical Alignment Drift

A growing body of evidence contradicts the null model’s predictions. Alignment degradation is not rare, and is, in fact, common, rapid and strongly dependent on task structure in ways not explained by random projection.

**Benign Tasks Break Safety.** The strongest challenge to the predictions of Theorem 4.1 comes from fine-tuning on datasets with no harmful content. [25] demonstrate that fine-tuning LLaMa-2-7B and GPT-3.5-Turbo on standard instruction datasets for fewer than 100 steps causes sharp drop off (within five gradient steps) in safety refusal on adversarial prompts, consistent with the results of [20]. Their work further shows that even carefully filtered data can trigger such jailbreaks with only 100 benign instruction-following samples. [6] and [44] further discuss this “emergent misalignment” by showing that narrow task adaptation systematically erodes safety boundaries.

**Non-Uniform Sensitivity and Threshold Effects.** If degradation arose from random overlap as Theorem 4.1 predicts, we would expect smooth, gradual decline proportional to training time. Instead, [60] observe sharp transitions, where modifying a specific 3% of weights causes catastrophic safety collapse, while random 3% samples have negligible effect. This non-uniformity directly contradicts the null model’s assumption that degradation scales with average curvature across all parameters. Moreover, [44] and [57] document threshold behaviors where models remain stable for several steps before rapid failure onset, again inconsistent with smooth probabilistic degradation.

**Task-Dependent Variation.** The random-update null model of Theorem 4.1 further implies that utility loss should be predictable from data overlap: tasks with distributions similar to  $\mathcal{D}_i$  should cause degradation (overlapping data would induce updates with non-negligible projection onto  $M_i$ ) while dissimilar tasks should be safe. However, [26] show that task influence is highly variable and poorly predicted by natural metrics. [25] formalize this by demonstrating that gradient-based selection finds harmful subsets within benign data—not by simple similarity-based computations.

All of these phenomena reveal a consistent signature for alignment degradation: it is systematic, it is rapid, and it is task-dependent (not similarity based).

### 4.3 Failure of Static Geometry

Our key observation is that the cause of the null model failing to explain observed empirical behaviors is its assumption that  $F_i(\theta)$  is static, i.e., the sensitivity subspace  $M_i$  remains fixed throughout fine-tuning. Such an assumption is, in fact, only valid in a vanishingly small neighborhood of  $\theta^*$ . In reality, both the magnitude and orientation of the Fisher information matrix evolve as parameters change along the trajectory  $\theta(t)$ . Thus, directions that were initially safe become dangerous. Moreover, fine-tuning updates themselves are highly structured in accordance with the task objective, and thus not random.

## 5 Alignment Instability Condition

Section 4 demonstrated that static, first-order analysis cannot explain empirical alignment degradation. The resolution requires understanding how fine-tuning trajectories evolve in a dynamic geometry where both update direction and the sensitive structure change over time.

We observe that in practice, fine-tuning induces a continuous trajectory  $\theta(t)$  through parameter space along which the Fisher information varies: its spectrum shifts, its eigenspaces rotate, and the alignment sensitivity subspace  $M_i$  itself evolves. As a result, directions initially orthogonal to  $M_i(\theta^*)$  need not remain, and small first-order projections can accumulate into significant degradation through higher-order effects. This section formalizes the precise structural conditions under which such accumulation becomes inevitable.

### 5.1 Intuition Behind Two Degradation Regimes

Let  $g(\theta) = \nabla L_{\text{FT}}(\theta)$  denote the fine-tuning gradient and  $M_i(\theta)$  the sensitivity subspace for skill  $S_i$  at parameters  $\theta$ . We distinguish two fundamentally different mechanisms of alignment degradation based on the initial overlap (measured by Euclidean projection) between the update direction and sensitive subspace.

**First-Order Regime.** If the initial fine-tuning gradient has substantial projection onto  $M_i$ , that is,  $\|P_i(\theta^*)g(\theta^*)\| \geq c$  for some non-negligible constant  $c > 0$ , then alignment degradation begins immediately. By the gradient flow  $\dot{\theta}(t) = -g(\theta(t))$  with  $\theta(0) = \theta^*$  we readily obtain that  $\|P_i(\theta^*)(\theta(t) - \theta^*)\| = \Theta(ct)$ . Combining with Proposition 3.3 yields quadratic degradation from first-order overlap. This regime corresponds to the more trivial fine-tuning tasks that directly modify alignment-relevant behaviors, requiring no additional geometric analysis beyond first-order projections. The remainder of this section focuses on the more subtle second-order regime.

**Second-Order Regime.** Consider instead the scenario where the initial gradient is nearly orthogonal to the sensitivity subspace:  $\|P_i(\theta^*)g(\theta^*)\| < \varepsilon$  for small  $\varepsilon > 0$ . Here, the first-order term vanishes and, naively, the trajectory should remain safe. However, this conclusion assumes that  $M_i$  and  $g$  remain fixed. When accounting for how such quantities evolve, second-order terms dominate. Specifically, expanding  $\theta(t)$  along the gradient flow to second order, we obtain:

$$\theta(t) - \theta^* = -tg(\theta^*) + \frac{t^2}{2} \nabla g(\theta^*)g(\theta^*) + O(t^3).$$

Under the projection, the first term contributes only  $\varepsilon t$ , whereas the second term—the directional derivative of the gradient—can have substantial projection onto  $M_i$ . This acceleration term captures how the curvature of the fine-tuning loss bends the trajectory.

## 5.2 Geometric Intuition: Curvature as Steering

If parameter space were naturally Euclidean and  $\mathbf{F}_i(\theta)$  were constant, gradient flow would proceed in straight lines and a trajectory starting perpendicular to  $M_i$  would remain perpendicular, preserving alignment. In reality, the information geometry of neural networks is non-Euclidean [2, 54], with distance between distributions depending upon the position. Crucially, as the model adapts through parameter space, the metric changes, the subspaces rotate, and the gradient curves.

**Curvature Coupling.** The key quantity,  $\nabla g(\theta^*)g(\theta^*)$ , is the directional derivative of the fine-tuning gradient. Geometrically, this is the acceleration of the trajectory, or, how much the direction of steepest descent changes. Most important for the present study, even if  $g(\theta^*)$  points away from  $M_i$ , the derivative  $\nabla g(\theta^*)$  can point toward it. Through fine-tuning, we then have three interacting mechanisms dictating the misalignment process:  $M_i$  must have high curvature so movement into it damages safe behaviors (this determines loss magnitude, analogous to the first-order regime), the orthogonality of a trajectory to this subspace dictates potential movement within this space (initial condition of the second-order regime), and the non-trivial coupling of  $\nabla g(\theta^*)g(\theta^*)$  can accelerate projection into this space (the acceleration mechanism unique to second-order regime).

## 5.3 Formal Definition

We now formalize the structural condition under which curvature forces trajectories into alignment-sensitive directions despite initial orthogonality. Throughout, for analytical clarity and without loss of generality, we assume that  $\mathbf{F}_i(\theta)$  and  $g(\theta)$  are twice continuously differentiable in a neighborhood of  $\theta^*$ .

**Definition 5.1** (Alignment Instability Condition). *Fix a skill  $S_i$ . Let  $\mathbf{F}_i(\theta^*)$  have eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq 0$ , and let  $M_i(\theta^*)$  be the span of the top  $d$ -eigenvectors of  $\mathbf{F}_i(\theta^*)$ , with  $P_i(\theta^*)$  the orthogonal projection onto  $M_i(\theta^*)$ . We say that  $\theta^*$  satisfies the **Alignment Instability Condition (AIC)** for skill  $S_i$  with parameters  $(d, \lambda, \gamma, \varepsilon)$  if the following hold:*

1. **Low-Rank Sensitivity:** *The Fisher spectrum at  $\theta^*$  is concentrated in a  $d$ -dimensional subspace:  $\sum_{j>d} \lambda_j \leq \varepsilon$  and  $\lambda_d \geq \lambda$ .*
2. **Initial Orthogonality:** *The fine-tuning update direction is initially nearly orthogonal to the sensitivity subspace:  $\|P_i(\theta^*)g(\theta^*)\| \leq \varepsilon$ .*
3. **Curvature Coupling:** *The curvature of the fine-tuning objective induces nontrivial acceleration into the sensitivity subspace:*

$$\|\mathbf{F}_i(\theta^*)^{1/2}P_i(\theta^*)\nabla g(\theta^*)g(\theta^*)\| \geq \gamma.$$

At a high level, the AIC identifies the regime where alignment appears stable initially (Condition 2) yet is structurally doomed (Condition 3) because it resides in a sharp, low-dimensional region (Condition 1). More formally, Condition 1 isolates the small set of directions where a skill  $S_i$  is fragile, with the bounds ensuring sufficient curvature concentrated in the first  $d$  eigenvectors and the tail contributing negligibly. Condition 2 captures the natural defense measure of projection away from sensitive subspaces, like that of [65]. However, we note that our analysis in Section 6 reveals this is a snapshot solution: unless projections are continuously updated to track the evolving sensitive subspace, curvature coupling (Condition 3) inevitably bends the trajectory back into sensitive regions over time. More specifically, this captures the rotational geometry crux: as we move along the trajectory, the gradient rotates toward  $M_i$ , with the parameter  $\gamma$  quantifying the strength of coupling.

## 6 Consequences of AIC

We now develop the main theoretical consequences of the AIC, establishing that alignment degradation under these conditions is not merely possible but structurally inevitable. These results enable both qualitative understanding (why benign tasks break safety) and quantitative prediction (how fast degradation occurs). Throughout this section, we fix a skill  $S_i$  and assume that  $\theta^*$  satisfies Definition 5.1 with parameters  $(d, \lambda, \gamma, \varepsilon)$ . Our approach proceeds in two steps, mirroring the geometric intuition from Section 5. We first quantify how much utility degrades when the trajectory enters the sensitive subspace  $M_i$ . Subsequently, we prove that the AIC forces the trajectory to enter  $M_i$

at a quadratic rate. Combining the two results, we derive the quartic scaling law. The key insight is that these are separate geometric phenomena that when composed yield the quartic scaling law.

For clarity of presentation, the arguments in this section are carried out with respect to the sensitive subspace defined by the Fisher information at  $\theta^*$ . We note, however, that the conclusions are robust to local variation of the Fisher information and to rotation of the sensitive subspace along the trajectory; accounting for these effects introduces only constant-factor and lower-order corrections and is deferred to Appendix B.

## 6.1 Projection of Misalignment Inevitability

We begin by formalizing the intuition that alignment resides in a small number of sharp directions. Movement into the subspaces  $M_i$  then degrades skill  $S_i$ , with loss growing quadratically in the projection magnitude.

Recall from Section 3 that utility loss locally takes the form  $\Delta u_i \approx \frac{1}{2} \Delta\theta^\top F_i(\theta^*) \Delta\theta$ . Here, the Fisher matrix characterizes the local curvature of alignment utility, identifying directions along which small parameter changes induce large distributional shifts. AIC Condition 1 concentrates the eigenspectrum, implying that movement in  $M_i$  causes this divergence to scale with  $\lambda$  whereas movement outside is negligible. Effectively, only the component projected into  $M_i$  matters for utility. The following theorem establishes the “cost” for entering sensitive regions.

**Theorem 6.1.** *Under Assumption 1 and assuming AIC holds, then there exists a neighborhood  $U$  of  $\theta^*$  such that for all  $\theta \in U$  with update step  $\Delta\theta = \theta - \theta^*$ :*

$$\Delta u_i(\theta) \geq \frac{1}{2} \|F_i^{1/2} P_i(\theta - \theta^*)\|^2 - O(\|\Delta\theta\|)^3.$$

Consequently, if the projection satisfies  $\|P_i(\Delta\theta)\| \geq \delta$  and the step size  $\|\Delta\theta\|$  is sufficiently small, we have  $\Delta u_i(\theta) = \Omega(\lambda\delta^2)$ .

The proof follows from decomposing  $F_i(\theta^*)$  and writing  $\Delta\theta$  in the eigenbasis. Projecting this change with  $P_i$  then decomposes into the overlapping and orthogonal components, both of which can be bounded according to the AIC conditions. Full details are presented in Appendix A. Observe that this theorem characterizes the loss if projection occurs, but does not capture whether such overlap occurs during fine-tuning. We now focus on showing that, under the AIC, drift into  $M_i$  is dynamically forced.

## 6.2 Curvature-Induced Drift

Even when the initial fine-tuning gradient is nearly orthogonal to  $M_i$ , the curvature of the loss landscape forces the trajectory to drift into the sensitive subspace at a quadratic rate. This drift is dictated not by the velocity  $(-\dot{g}(\theta^*))$ , but the acceleration  $(\nabla g(\theta^*)g(\theta^*))$  which monitors how the direction of steepest descent changes along a trajectory. AIC Condition 3 bounds this factor, ensuring that projection  $M_i$  increases through training. Thus, our initially small overlap trajectory drifts into sensitive regions. Note that this acceleration is due to the curvature of the fine-tuning objective. Gradient descent follows the local geometry with no mechanism to anticipate or counteract curvature-induced drift.

**Theorem 6.2.** *Fix a skill  $S_i$ . Let  $g(\theta) := \nabla L_{FT}(\theta)$  be the fine-tuning gradient and consider the gradient-flow trajectory defined by  $\dot{\theta}(t) = -g(\theta(t))$  and  $\theta(0) = \theta^*$ . Assuming  $g$  is  $C^2$  in an  $r$ -ball of  $\theta^*$  and the AIC holds, then there exists a  $t_0 > 0$  such that for  $t \in (0, t_0]$  we have that*

$$\|F_i^{1/2} P_i(\theta(t) - \theta^*)\| \geq \frac{\gamma}{2} t^2 - \varepsilon t - O(t^3).$$

The proof follows from Taylor expanding the trajectory along the gradient flow to second order. Projecting this trajectory onto  $M_i$  via  $P_i$ , we can then bound the resulting terms using the AIC in combination with the reverse triangle inequality to complete the result. Full details are in Appendix A.

## 6.3 Combing Geometry and Dynamics

Theorems 6.1 and 6.2 address complementary aspects of alignment degradation. Composing these results yields our main prediction: alignment utility degrades as the fourth power of training evolution.

**Corollary 6.3** (Quartic Onset of Alignment Degradation). *Fix a skill  $S_i$  and let  $\theta(t)$  be a fine-tuning trajectory satisfying the assumptions of Theorems 6.1 and 6.2. Then there exists  $t_0 > 0$  such that for all  $t \in [0, t_0]$ , the utility loss for skill  $S_i$  satisfies  $\Delta u_i = \Omega(\lambda \gamma^2 t^4)$ .*

The quartic law provides a quantitative explanation for rapid safety collapse observed in practice [25, 44]. Thus, our results show that this perceived sudden failure is actually a smooth quartic due to curvature coupling.

We lastly note that the preceding results are stated with respect to the sensitive subspace  $M_i$  defined by the Fisher information at  $\theta^*$ . In practice,  $F_i(\theta)$  varies along the fine-tuning trajectory, inducing rotation of the corresponding eigenspaces. Under mild local regularity, this rotation can be controlled via standard Davis-Kahan perturbation bounds, and the proofs extend with only constant-factor and lower-order corrections to the drift and loss bounds. We provide the formal statement and proof in Appendix B.

## 7 Experiments

Our theory establishes that alignment degradation is geometrically inevitable under the AIC, with the Fisher information matrix (FIM) characterizing the curvature structure that governs this fragility. We now validate (1) that the FIM exhibits low-rank structure required by the AIC, and (2) that our notion of geometric overlap successfully predicts which fine-tuning tasks will degrade alignment.

### 7.1 Low-Rank Fisher Structure

Our theory operates in weight space, where the FIM’s eigenvector define alignment-sensitive directions (Definition 3.4). However, prior empirical work characterizes alignment through low-rank structure in representation space (activations), rather than weight space [3, 60, 65, 66]. Specifically, these works show that for a linear layer  $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ , the average gradient  $\nabla L_W(T)$  with respect to alignment sequences  $T$  typically concentrates in an  $r$ -dimensional subspace of the representation space, with  $r \ll \min\{d_{\text{out}}, d_{\text{in}}\}$ .

This distinction is crucial because representation-space low-rank identifies which neurons are sensitive, while weight-space low-rank via FIM quantifies how much curvature exists along sensitive directions ( $\lambda$  in our theory). These are complementary perspectives on the same underlying geometry, but our theory relies on the FIM formulation to characterize second-order drift (Theorem 6.2). In Appendix A.4, we prove that these two notions are related: if  $\nabla L_W(T)$  concentrates in an  $r$ -dimensional representation space, then the FIM inherits low-rank, establishing AIC Property 1. Thus, existing empirical evidence implies our weight-space Fisher structure, but we validate this directly.

**Computational Approach.** Computing the full FIM for billion-parameter models is intractable. Following [66], we use block-wise analysis, treating each module separately. For a weight matrix  $W \in \mathbb{R}^{d \times d}$ , we first flatten to  $\mathbb{R}^{d^2}$ , then project to a low-dimensional space using Rademacher random projection [25, 30, 41, 61]. We then compute gradients on 100 safe completions from BeaverTails [28] and form the projected FIM.

**Results.** In Figure 2, we visualize sorted eigenvalues for  $W_{\text{up}}$ ,  $W_{\text{gate}}$  and  $W_{\text{down}}$  across all MLP blocks (with the remaining modules deferred to Appendix D.1). Notably, the eigenvalue decay exhibits the sharp low-rank structure leveraged by our theory. This validates AIC Property 1 empirically and confirms that the connection from representation-space to weight-space low-rank holds in practice.

### 7.2 Geometric Overlap and Misalignment

Recent work documents emergent misalignment (EM): fine-tuning on seemingly benign datasets causes safety degradation [6, 57]. [59] explain this through “persona shifts” identified via Sparse Autoencoders (SAEs) [8, 15, 21] in representation space, a diagnostic approach that requires access to a target evaluation set to identify which personas are problematic. We demonstrate that our geometric framework offers a complementary and predictive explanation. Specifically, EM arises from geometric overlap between fine-tuning updates and the Fisher-defined alignment subspace, quantifiable *before* observing safety failures.

We approximate the utility loss from Proposition 3.3 using an **Overlap Score (OS)**. For each fine-tuned model, we extract per-module weight changes  $\Delta W \in \mathbb{R}^{d \times d}$ , flatten to  $\mathbb{R}^{d^2}$ , apply the same Rademacher projection  $P_{\text{random}} \in$

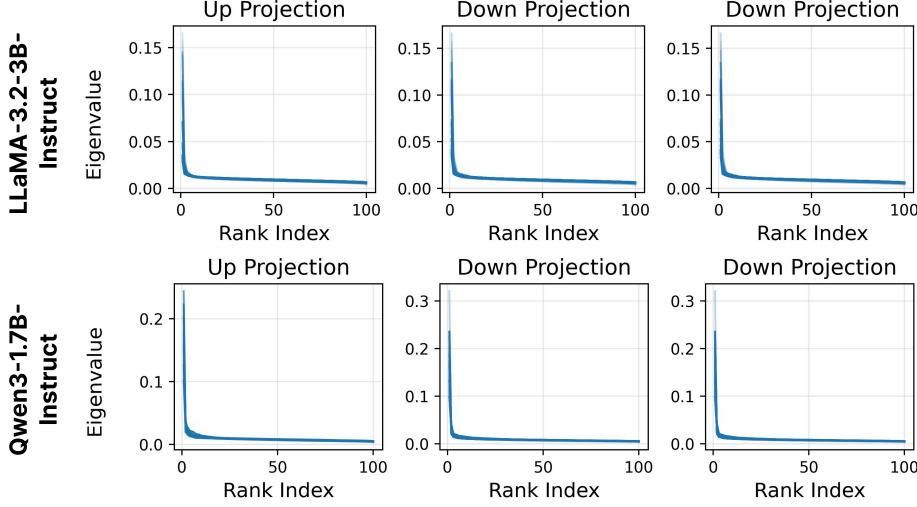


Figure 2: Top eigenvalues of FIM approximated over 100 random samples from *BeaverTail*'s safe subset. Each subplot consists of multiple lines in different transparency levels for different layer indices, with each layer showing a similar low-rank structure.

$\mathbb{R}^{4096 \times d^2}$  used for the FIM, and compute:

$$\mathbf{OS} = (\mathbf{P}_{\text{random}} \text{vec}(\Delta W))^T \mathbf{F} (\mathbf{P}_{\text{random}} \text{vec}(\Delta W)).$$

This measures how much the fine-tuning update projects onto the high-curvature alignment subspace. Higher OS predicts greater alignment degradation, enabling prospective risk assessment.

**Experimental Setup.** We fine-tune Qwen3-1.7B-Instruct [53] and LLaMA-3.2-3B-Instruct [23] using both LoRA [27] and full fine-tuning on datasets in three categories: (1) *Benign* (SamSum, Alpaca Random 100, GSM8K Random 100), (2) *Seemingly Benign* (Alpaca Top 100 via gradient matching [25], Risky Financial Advice, Bad Medical Advice), and (3) *Harmful* (Pure Bad [44]). Following standard protocols [25, 58, 65], we evaluate using Gemini-2.5-Flash [13] as a judge to assign Harmfulness Scores (HS, 1-5 scale) on 520 AdvBench queries [67]. Full dataset and training details in Appendix D.

Table 1 showcases that the three dataset categories produce distinct harmfulness levels, with *Seemingly Benign* datasets achieving HS scores comparable to explicitly harmful data. Critically, all fine-tuning increases harmfulness relative to the base model, but the magnitude is strongly modulated by dataset characteristics. This confirms our framework’s prediction that alignment degradation is unavoidable but scales with task-geometry coupling. In full fine-tuning, our Overlap Score successfully distinguishes dangerous from safe tasks (Figure 3). The *Seemingly Benign* datasets exhibit OS patterns similar to *Pure Bad*, while truly benign datasets show substantially lower overlap. This validates that geometric coupling plays a role in degradation: datasets with no apparent overlap with harmful content still break safety when they couple strongly to alignment-sensitive parameters. One exception is *Alpaca Top 100*, where OS does not clearly separate from benign datasets, likely due to (1) noise in random projection to lower dimensions, or (2) Alpaca’s diversity compared to domain-specific datasets.

**Limitations.** We additionally note that LoRA does not show the same clear OS-to-HS correlation (Figure 3, right panels). We hypothesize this occurs because LoRA introduces “intruder dimensions” [50] that shift the model’s top eigenvectors, making overlap with the pre-training Fisher less reliable. This supports our theoretical result on parametrization dependence. Additionally, the first-order OS may be insufficient for LoRA because its constrained update structure amplifies the induced coupling condition (AIC Condition 3), requiring second-order curvature analysis via  $\nabla g(\theta^*)$ . Computing this remains computationally prohibitive [33]; developing tractable curvature estimation is critical future work.

In spite of this limitation, we note that while [59]’s SAE-based approach identifies what changes in representations, our geometric framework explains *why* these changes occur (curvature forces drift into  $\mathcal{M}_t$ ) and predicts which tasks will

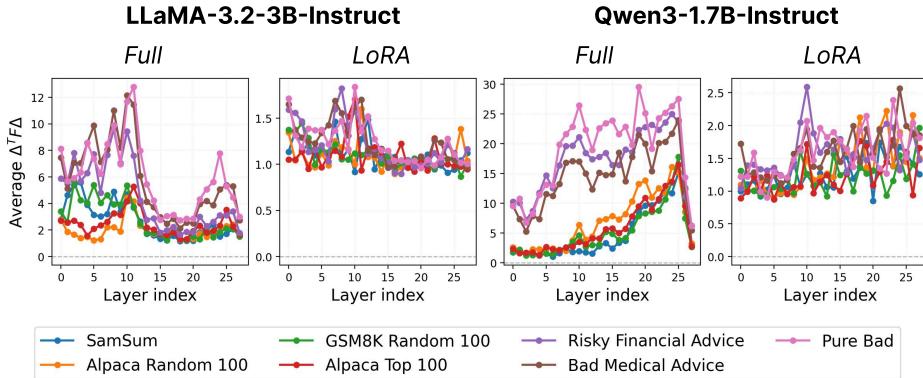


Figure 3: Average Overlap Score per Transformer Block of 7 Fine-Tuning Datasets. Each value represents the average OS of all modules in the block.

cause them via OS. Critically, our approach requires only the fine-tuning dataset and pre-computed FIM—no access to target evaluation sets or trained SAEs—making it applicable for prospective risk assessment before deployment.

Table 1: Harmfulness Score (HS) across fine-tuning datasets

Finetuning Dataset	Qwen3-1.7B-Instruct		LLaMA-3.2-3B-Instruct	
	LoRA	Full	LoRA	Full
No Fine-tuning		1.24		1.11
<b>Benign</b>				
SamSum	2.07	2.15	1.10	1.77
Alpaca Random 100	2.68	2.92	1.31	1.76
GSM8K Random 100	3.63	3.42	1.10	1.38
<b>Seemingly Benign</b>				
Alpaca Top 100	3.15	3.68	1.18	3.70
Risky Financial Advice	3.94	3.78	4.71	3.82
Bad Medical Advice	3.92	3.86	2.46	3.65
<b>Harmful</b>				
Pure Bad	4.55	3.98	4.89	4.51

## 8 Discussion & Implications

We have established that alignment degradation is an intrinsic geometric property of gradient-based fine-tuning, not an artifact of adversarial attacks or poor data curation. The AIC reveals why first-order defenses fail: while projecting updates into null spaces provides temporary protection, second-order curvature effects systematically bend trajectories back into sensitive subspaces. Our quartic scaling law explains the rapid onset of safety failures observed empirically and demonstrates that current defenses are fundamentally insufficient because they ignore the dynamic, curved geometry of the parameter space. Moreover, our framework enables practical diagnostic tools for safer fine-tuning. We propose that practitioners can estimate the coupling parameter  $\gamma$  before training to assess risk, monitor the projection size during training for early warning signals, and audit these values after training to quantify alignment degradation. More fundamentally, effective preservation of alignment requires curvature-aware methods that track evolving sensitive subspaces, constrain second-order acceleration into these regions, and balance competing curvature constraints across multiple alignment skills.

Our results challenge the assumption that alignment preservation is primarily an engineering problem solvable

through better data filtering. Instead, it is a fundamental question of geometric control in high-dimensional curved spaces, with profound implications for open-weight model deployment. The structural fragility we expose cannot be eliminated through pre-deployment alignment alone, requiring the community to reconsider whether current deployment paradigms are compatible with safety requirements.

## References

- [1] ALKAED, M., ABIOYE, S., QAYYUM, A., MEKKI, Y. M., BERROU, I., ABDALLAH, M., AL-FUQAH, A., BILAL, M., AND QADIR, J. Open foundation models in healthcare: Challenges, paradoxes, and opportunities with genai driven personalized prescription. *arXiv preprint arXiv:2502.04356* (2025).
- [2] AMARI, S.-I. Information geometry of neural networks—an overview—. *Mathematics of Neural Networks: Models, Algorithms and Applications* (1997), 15–23.
- [3] ARDITI, A., OBESO, O., SYED, A., PALEKA, D., PANICKSSERY, N., GURNEE, W., AND NANDA, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems 37* (2024), 136037–136083.
- [4] ARORA, S., AND GOYAL, A. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936* (2023).
- [5] ARTURI, D. A. R., ZHANG, E., ANSAH, A. A., ZHU, K., PANDA, A., AND BALWANI, A. Shared parameter subspaces and cross-task linearity in emergently misaligned behaviour. In *Mechanistic Interpretability Workshop at NeurIPS 2025* (2025).
- [6] BETLEY, J., WARNCKE, N., SZTYBER-BETLEY, A., TAN, D., BAO, X., SOTO, M., SRIVASTAVA, M., LABENZ, N., AND EVANS, O. Training large language models on narrow tasks can lead to broad misalignment. *Nature* 649, 8097 (2026), 584–589.
- [7] BHANDARI, K. R., CHEN, P.-Y., AND GAO, J. Forecasting open-weight AI model growth on huggingface. *arXiv preprint arXiv:2502.15987* (2025).
- [8] BRICKEN, T., TEMPLETON, A., BATSON, J., CHEN, B., JERMYN, A., CONERLY, T., TURNER, N., ANIL, C., DENISON, C., ASKELL, A., LASENBY, R., WU, Y., KRAVEC, S., SCHIEFER, N., MAXWELL, T., JOSEPH, N., HATFIELD-DODDS, Z., TAMKIN, A., NGUYEN, K., MCLEAN, B., BURKE, J. E., HUME, T., CARTER, S., HENIGHAN, T., AND OLAH, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread* (2023). <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- [9] BURTON, J. W., LOPEZ-LOPEZ, E., HECHTLINGER, S., RAHWAN, Z., AESCHBACH, S., BAKKER, M. A., BECKER, J. A., BERDITCHEVSKAIA, A., BERGER, J., BRINKMANN, L., ET AL. How large language models can reshape collective intelligence. *Nature human behaviour* 8, 9 (2024), 1643–1655.
- [10] CHOKSI, M. Z., MANDEL, I., AND BENTHALL, S. The brief and wondrous life of open models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025), pp. 3224–3240.
- [11] CHUA, K., LEI, Q., AND LEE, J. D. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems 34* (2021), 8871–8884.
- [12] COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M., TWOREK, J., HILTON, J., NAKANO, R., ET AL. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- [13] COMANICI, G., BIEBER, E., SCHAEKERMANN, M., PASUPAT, I., SACHDEVA, N., DHILLON, I., BLISTEIN, M., RAM, O., ZHANG, D., ROSEN, E., ET AL. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
- [14] COVER, T. M. *Elements of information theory*. John Wiley & Sons, 1999.
- [15] CUNNINGHAM, H., EWART, A., RIGGS, L., HUBEN, R., AND SHARKEY, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600* (2023).

- [16] DAS, A., BORAH, A., JAIN, V., AND CHADHA, A. Alignguard-lora: Alignment-preserving fine-tuning via fisher-guided decomposition and riemannian-geodesic collision regularization. *arXiv preprint arXiv:2508.02079* (2025).
- [17] DAVIS, C., AND KAHAN, W. M. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis* 7, 1 (1970), 1–46.
- [18] DONG, Y., MU, R., ZHANG, Y., SUN, S., ZHANG, T., WU, C., JIN, G., QI, Y., HU, J., MENG, J., ET AL. Safeguarding large language models: A survey. *Artificial Intelligence Review* 58, 12 (2025), 382.
- [19] DU, S. S., HU, W., KAKADE, S. M., LEE, J. D., AND LEI, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434* (2020).
- [20] FRASER, K. C., DAWKINS, H., NEJADGHOOLI, I., AND KIRITCHENKO, S. Fine-tuning lowers safety and disrupts evaluation consistency. *arXiv preprint arXiv:2506.17209* (2025).
- [21] GAO, L., LA TOUR, T. D., TILLMAN, H., GOH, G., TROLL, R., RADFORD, A., SUTSKEVER, I., LEIKE, J., AND WU, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations* (2025).
- [22] GLIWA, B., MOCHOL, I., BIESEK, M., AND WAWER, A. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 70–79.
- [23] GRATTAFIORI, A., DUBEY, A., JAURHI, A., PANDEY, A., KADIAN, A., AL-DAHLE, A., LETMAN, A., MATHUR, A., SCHELTEN, A., VAUGHAN, A., ET AL. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [24] GUAN, Z., HU, M., ZHU, R., LI, S., AND VULLIKANTI, A. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. In *Forty-second International Conference on Machine Learning* (2025).
- [25] HE, L., XIA, M., AND HENDERSON, P. What is in your safe data? Identifying benign data that breaks safety. In *First Conference on Language Modeling* (2024).
- [26] HSIUNG, L., PANG, T., TANG, Y.-C., SONG, L., HO, T.-Y., CHEN, P.-Y., AND YANG, Y. Your task may vary: A systematic understanding of alignment and safety degradation when fine-tuning LLMs.
- [27] HU, E. J., SHEN, Y., WALLIS, P., ALLEN-ZHU, Z., LI, Y., WANG, S., WANG, L., CHEN, W., ET AL. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
- [28] JI, J., LIU, M., DAI, J., PAN, X., ZHANG, C., BIAN, C., CHEN, B., SUN, R., WANG, Y., AND YANG, Y. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems* 36 (2023), 24678–24704.
- [29] JIANG, X., WANG, X., AND STICH, S. U. Loram: Low-rank adaptation of large language models on manifold. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*.
- [30] JOHNSON, W. B., LINDENSTRAUSS, J., ET AL. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics* 26, 189–206 (1984), 1.
- [31] KAPOOR, S., BOMMASANI, R., KLYMAN, K., LONGPRE, S., RAMASWAMI, A., CIHON, P., HOPKINS, A., BANKSTON, K., BIDERMAN, S., BOGEN, M., ET AL. On the societal impact of open foundation models. *arXiv preprint arXiv:2403.07918* (2024).
- [32] KAUSHIK, P., CHAUDHARI, S., VAIDYA, A., CHELLAPPA, R., AND YUIILLE, A. The universal weight subspace hypothesis, 2025.
- [33] KUNSTNER, F., HENNIG, P., AND BALLES, L. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems* 32 (2019).
- [34] LEE, A., BAI, X., PRES, I., WATTENBERG, M., KUMMERFELD, J. K., AND MIHALCEA, R. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Forty-first International Conference on Machine Learning* (2024).

- [35] LI, C., FARKHOOR, H., LIU, R., AND YOSINSKI, J. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838* (2018).
- [36] LIU, Z., NIE, Y., TAN, Y., YUE, X., CUI, Q., WANG, C., ZHU, X., AND ZHENG, B. Safety alignment for vision language models. *arXiv preprint arXiv:2405.13581* (2024).
- [37] MALLADI, S., WETTIG, A., YU, D., CHEN, D., AND ARORA, S. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning* (2023), PMLR, pp. 23610–23641.
- [38] MANTRI, K. S. I., ELIASOF, M., SCHÖNLIEB, C.-B., AND RIBEIRO, B. Rethinking fine-tuning through geometric perspective. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models* (2024).
- [39] NIU, Y., XIAO, H., LIU, D., CHEN, N., AND LI, J. Mitigating the safety alignment tax with null-space constrained policy optimization. *arXiv preprint arXiv:2512.11391* (2025).
- [40] OSBORNE, C., DING, J., AND KIRK, H. R. The ai community building the future? a quantitative analysis of development activity on hugging face hub. *Journal of Computational Social Science* 7, 2 (2024), 2067–2105.
- [41] PARK, S. M., GEORGIEV, K., ILYAS, A., LECLERC, G., AND MKADRY, A. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning* (2023), ICML ’23, JMLR.org.
- [42] PENG, S. Y., CHEN, P.-Y., HULL, M., AND CHAU, D. H. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems* 37 (2024), 95692–95715.
- [43] QI, X., WEI, B., CARLINI, N., HUANG, Y., XIE, T., HE, L., JAGIELSKI, M., NASR, M., MITTAL, P., AND HENDERSON, P. On evaluating the durability of safeguards for open-weight LLMs. *arXiv preprint arXiv:2412.07097* (2024).
- [44] QI, X., ZENG, Y., XIE, T., CHEN, P.-Y., JIA, R., MITTAL, P., AND HENDERSON, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693* (2023).
- [45] RAJABI, S., AND RAMBHATLA, S. Enhancing fine-tuning efficiency of LLMs through gradient subspace tracking. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning* (2024).
- [46] RIFAI, S., DAUPHIN, Y. N., VINCENT, P., BENGIO, Y., AND MULLER, X. The manifold tangent classifier. *Advances in neural information processing systems* 24 (2011).
- [47] SAGUN, L., EVCİ, U., GUNEV, V. U., DAUPHIN, Y., AND BOTTOU, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454* (2017).
- [48] SCHULMAN, J., LEVINE, S., ABBEEL, P., JORDAN, M., AND MORITZ, P. Trust region policy optimization. In *International conference on machine learning* (2015), PMLR, pp. 1889–1897.
- [49] SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A., AND KLIMOV, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [50] SHUTTLEWORTH, R. S., ANDREAS, J., TORRALBA, A., AND SHARMA, P. LoRA vs full fine-tuning: An illusion of equivalence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025).
- [51] SOLIGO, A., TURNER, E., RAJAMANOHARAN, S., AND NANDA, N. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618* (2025).
- [52] TAORI, R., GULRAJANI, I., ZHANG, T., DUBOIS, Y., LI, X., GUESTRIN, C., LIANG, P., AND HASHIMOTO, T. B. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [53] TEAM, Q. Qwen3 technical report, 2025.
- [54] THIRUTHUMMAL, A. A., KIM, E.-J., AND SHELYAG, S. Information geometry of evolution of neural network parameters while training. *Neurocomputing* 597 (2024), 128007.
- [55] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., ET AL. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

- [56] TRIPURANENI, N., JIN, C., AND JORDAN, M. Provable meta-learning of linear representations. In *International conference on machine learning* (2021), PMLR, pp. 10434–10443.
- [57] TURNER, E., SOLIGO, A., TAYLOR, M., RAJAMANOHARAN, S., AND NANDA, N. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613* (2025).
- [58] WANG, J., LI, J., LI, Y., QI, X., HU, J., LI, Y., McDANIEL, P., CHEN, M., LI, B., AND XIAO, C. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. *arXiv preprint arXiv:2402.14968* (2024).
- [59] WANG, M., LA TOUR, T. D., WATKINS, O., MAKELOV, A., CHI, R. A., MISERENDINO, S., WANG, J., RAJARAM, A., HEIDECKE, J., PATWARDHAN, T., ET AL. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823* (2025).
- [60] WEI, B., HUANG, K., HUANG, Y., XIE, T., QI, X., XIA, M., MITTAL, P., WANG, M., AND HENDERSON, P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162* (2024).
- [61] XIA, M., MALLADI, S., GURURANGAN, S., ARORA, S., AND CHEN, D. Less: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning* (2024), ICML’24, JMLR.org.
- [62] YANG, J. Q., SALAMATIAN, S., SUN, Z., SURESH, A. T., AND BEIRAMI, A. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)* (2024), IEEE, pp. 2027–2032.
- [63] YANG, K., TAO, G., CHEN, X., AND XU, J. Alleviating the fear of losing alignment in llm fine-tuning. In *2025 IEEE Symposium on Security and Privacy (SP)* (2025), IEEE, pp. 2152–2170.
- [64] YANG, Y., JIN, Q., HUANG, F., AND LU, Z. Adversarial prompt and fine-tuning attacks threaten medical large language models. *Nature Communications* 16, 1 (2025), 9011.
- [65] ZHANG, B., YANG, Y., GUO, D., GU, J., TORR, P., GHANEM, B., ET AL. A guardrail for safety preservation: When safety-sensitive subspace meets harmful-resistant null-space. *arXiv preprint arXiv:2510.14301* (2025).
- [66] ZHANG, J., HU, Y., CHEN, K., HE, L., MA, J., LOU, J., LI, D., LIU, J., YANG, X., AND JIA, R. Understanding and preserving safety in fine-tuned LLMs. *arXiv preprint arXiv:2601.10141* (2026).
- [67] ZOU, A., WANG, Z., CARLINI, N., NASR, M., KOLTER, J. Z., AND FREDRIKSON, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).

## A Omitted Proofs

### A.1 Proofs from Section 3

*Proof of Lemma 3.2.* Recall the definition of alignment utility for skill  $S_i$  is the expected log-likelihood under ideal distribution  $\mathcal{D}_i$ :

$$u_i(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\log \pi_\theta(y|x)]$$

The alignment loss is defined as  $\Delta u_i(\theta) = u_i(\theta^*) - u_i(\theta)$ . Substituting definitions, we obtain

$$\Delta u_i = \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\log \pi_{\theta^*}(y|x)] - \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\log \pi_\theta(y|x)]$$

By linearity of expectations we combine terms:

$$\Delta u_i = \mathbb{E}_{x \sim \mathcal{D}_i} \left[ \sum_y \mathcal{D}_i(y|x) \log \left( \frac{\pi_{\theta^*}(y|x)}{\pi_\theta(y|x)} \right) \right]$$

Invoking Assumption 1, we have that the aligned policy matches the ideal distribution,  $\pi_{\theta^*}(y|x) = \mathcal{D}_i(y|x)$  for all  $x \in \text{supp}(\mathcal{D}_i)$ . Substituting  $\mathcal{D}_i$  with  $\pi_{\theta^*}$ , we obtain

$$\Delta u_i = \mathbb{E}_{x \sim \mathcal{D}_i} \left[ \sum_y \pi_{\theta^*}(y|x) \log \left( \frac{\pi_{\theta^*}(y|x)}{\pi_\theta(y|x)} \right) \right]$$

where the term inside the expectation is precisely the KL divergence between  $\pi_{\theta^*}$  and  $\pi_\theta$ .  $\square$

*Proof of Proposition 3.3.* This follows from the second-order Taylor expansion of the KL divergence. The KL divergence,  $D_{\text{KL}}(\pi_{\theta^*} || \pi_\theta)$  is a function of  $\theta$  that is minimized when  $\theta = \theta^*$ . The gradient of the function at the minimum is zero and the Hessian is the Fisher information matrix. Expanding and taking the expectation over  $x \sim \mathcal{D}_i$  gives the desired result.  $\square$

### A.2 Proofs from Section 4

We first prove the isotropy of the uniform sphere for completeness.

**Lemma A.1.** *Let  $g \sim \text{Unif}(\mathbb{S}^{n-1})$ . Then  $\mathbb{E}_g[gg^\top] = \frac{1}{n}I_n$*

*Proof.* Let  $M := \mathbb{E}_g[gg^\top]$ . For any orthogonal matrix  $Q$ ,  $Qg$  has the same distribution as  $g$  by rotational invariance of  $\text{Unif}(\mathbb{S}^{n-1})$ . Hence,

$$\begin{aligned} M &= \mathbb{E}_g[gg^\top] \\ &= \mathbb{E}_g[(Qg)(Qg)^\top] \\ &= Q\mathbb{E}_g[gg^\top]Q^\top \\ &= QMQ^\top \end{aligned}$$

for any such  $Q$ . Thus,  $M$  commutes with orthogonal matrices. The only matrices with this property are scalar multiples of the identity,  $M = cI_n$  for some  $c \in \mathbb{R}$ . Taking traces,

$$\text{Tr}(M) = \mathbb{E}_g[\text{Tr}(gg^\top)] = \mathbb{E}_g[\|g\|^2] = 1$$

while  $\text{Tr}(cI_n) = cn$ . Therefore,  $c = 1/n$ , proving the claim.  $\square$

Moreover, since the Fisher information is a symmetric matrix, we have the following expected form of the quadratic term in our utility function.

**Lemma A.2.** *Let  $g \sim \text{Unif}(\mathbb{S}^{n-1})$  and let  $A \in \mathbb{R}^{n \times n}$  be symmetric. Then  $\mathbb{E}_g[gAg^\top] = \frac{1}{n}\text{Tr}(A)$ .*

*Proof.* Using  $\text{Tr}(uv^\top) = v^\top u$  and the cyclic nature of trace, we have  $g^\top Ag = \text{Tr}(g^\top Ag) = \text{Tr}(Agg^\top)$ . Taking expectation and applying the linearity of trace and expectation, we obtain  $\mathbb{E}_g[g^\top Ag] = \text{Tr}(A\mathbb{E}[gg^\top])$  which by Lemma A.1 finishes the proof.  $\square$

We now revisit the equation for change in utility, substituting  $\Delta\theta = \eta g$ . After taking the expectation and applying Lemma A.2 with  $A = F_i(\theta^*)$  we obtain the first bound of the proposition.

We proceed to verify the second claim. First, we prove a standard concentration inequality result on the rank- $d$  projector.

**Lemma A.3.** *Let  $g \sim \text{Unif}(\mathbb{S}^{n-1})$  and let  $P$  be any orthogonal projector of rank  $d$ . Then for any  $\varepsilon \in (0, 1)$ ,*

$$\Pr \left[ \left| \|Pg\|^2 - \frac{d}{n} \right| \geq \varepsilon \cdot \frac{d}{n} \right] \leq 2 \exp(-c\varepsilon^2 d)$$

for constant  $c > 0$ . In particular, with probability at least  $1 - 2e^{-ce^2 d}$ ,

$$(1 - \varepsilon) \frac{d}{n} \leq \|Pg\|^2 \leq (1 + \varepsilon) \frac{d}{n},$$

or  $\|Pg\|^2 = \Theta(d/n)$ .

*Proof.* By rotational invariance of  $g$ , the distribution of  $\|Pg\|^2$  depends only on the rank of  $P$ . Thus, we may assume  $P = \text{diag}(I_d, 0)$  without loss of generality. Let  $z \sim \mathcal{N}(0, I_n)$  and set  $g = z/\|z\|$  (ie. a construction of  $\text{Unif}(\mathbb{S}^{n-1})$ ). Define  $X := \sum_{j=1}^d z_j^2 \sim \chi_d^2$  and  $Y = \sum_{j=d+1}^n z_j^2 \sim \chi_{n-d}^2$ , with  $X$  and  $Y$  independent, and  $\|z\|^2 = X + Y$ . Then  $\|Pg\|^2 = \frac{\|Pz\|^2}{\|z\|^2} = \frac{X}{X+Y}$ . We can now invoke standard chi-square Chernoff bounds: for  $W \sim \chi_k^2$  and  $\varepsilon \in (0, 1)$ ,

$$\Pr [|W - k| \geq \varepsilon k] \leq 2 \exp(-c_0 \varepsilon^2 k)$$

for a constant  $c_0 > 0$ . Apply this to  $X$  with  $k = d$  and to  $X + Y \sim \chi_n^2$  with  $k = n$  to get that, with probability at least  $1 - 2e^{-c_0 \varepsilon^2 d} - 2e^{-c_0 \varepsilon^2 n} \geq 1 - 4e^{-c_0 \varepsilon^2 d}$ . This simultaneously yields  $(1 - \varepsilon)d \leq X \leq (1 + \varepsilon)d$  and  $(1 - \varepsilon)n \leq X + Y \leq (1 + \varepsilon)n$ . Combining these results we have

$$\frac{(1 - \varepsilon)d}{(1 + \varepsilon)n} \leq \frac{X}{X + Y} \leq \frac{(1 + \varepsilon)d}{(1 - \varepsilon)n}$$

for  $\varepsilon \in (0, 1/2]$  this further becomes

$$(1 - 2\varepsilon) \frac{d}{n} \leq \|Pg\|^2 \leq (1 + \varepsilon) \frac{d}{n}$$

which gives the ultimate result.  $\square$

We lastly need to bound the Fisher induced norm by  $\|Pg\|^2$ .

**Lemma A.4.** *Let  $F \succeq 0$  be symmetric with eigenvalues  $\lambda_1 \geq \dots \geq \lambda_n \geq 0$  and let  $P$  be the orthogonal projector onto the span of the top  $d$  eigenvectors of  $F$ . Then for any unit vector  $g$ ,*

$$g^\top F g \leq \lambda_1 \|Pg\|^2 + \lambda_{d+1}(1 - \|Pg\|^2)$$

In particular, we can further obtain

$$g^\top F g \leq \lambda_1 \|Pg\|^2 + \lambda_1(1 - \|Pg\|^2) = \lambda_{\max}(F).$$

*Proof.* Let  $V_d$  be the top- $d$  eigenspace of  $F$ , and write  $g = g_{\parallel} + g_{\perp}$  where  $g_{\parallel} := Pg \in V_d$  and  $g_{\perp} := (I - P)g \in V_d^{\perp}$ . Since  $V_d$  and  $V_d^{\perp}$  are orthogonal and  $F$  is diagonalizable in an orthonormal basis, we have

$$g^\top F g = g_{\parallel}^\top F g_{\parallel} + g_{\perp}^\top F g_{\perp}$$

as the cross terms equate to zero. By the Rayleigh quotient bound,  $g_{\parallel}^T F g_{\parallel} \leq \lambda_1 \|g_{\parallel}\|^2 = \lambda_1 \|Pg\|^2$ . On  $V_d^{\perp}$ , the largest eigenvalue is  $\lambda_{d+1}$ , thus by the same technique  $g_{\perp}^T F g_{\perp} \leq \lambda_{d+1}(1 - \|Pg\|^2)$  using  $\|g\| = 1$  and orthogonality. Summation yields

$$g^T F g \leq \lambda_1 \|Pg\|^2 + \lambda_{d+1}(1 - \|Pg\|^2)$$

with the final inequality following from  $(1 - \|Pg\|^2) \leq 1$ .  $\square$

We now collect the lemmas and prove Theorem 4.1

*Proof of Theorem 4.1.* Fix a skill  $S_i$  and write  $F := F_i(\theta^*) \succeq 0$ . Let  $\Delta\theta = \eta g$  where  $g \sim \text{Unif}(\mathbb{S}^{n-1})$  and  $\eta > 0$  is sufficiently small so that the local quadratic approximation applies. By the second-order expansion of the skill utility (cf. Lemma B.1),

$$\Delta u_i = \frac{1}{2} \Delta\theta^T F \Delta\theta + O(\|\Delta\theta\|^3) = \frac{\eta^2}{2} g^T F g + O(\eta^3), \quad (2)$$

since  $\|g\| = 1$ .

We begin with proving the expectation result. By Lemma B.2, for any symmetric matrix  $F$ ,

$$\mathbb{E}[g^T F g] = \frac{1}{n} \text{Tr}(F). \quad (3)$$

Taking expectations yields

$$\mathbb{E}[\Delta u_i] = \frac{\eta^2}{2n} \text{Tr}(F_i(\theta^*)) + O(\eta^3), \quad (4)$$

establishing part (1) of the theorem.

We now convert this to a high-probability result. Let  $P$  denote the orthogonal projector onto the span of the top  $d$  eigenvectors of  $F$ , and let  $\lambda_1 \geq \dots \geq \lambda_n \geq 0$  be the eigenvalues of  $F$ . By Lemma B.3, for any  $\varepsilon \in (0, 1)$ ,

$$\Pr\left[\|Pg\|^2 \leq (1 + \varepsilon)\frac{d}{n}\right] \geq 1 - 2e^{-c\varepsilon^2 d}. \quad (5)$$

Condition on this event. Lemma B.4 then implies that for any unit vector  $g$ ,

$$g^T F g \leq \lambda_1 \|Pg\|^2 + \lambda_{d+1}(1 - \|Pg\|^2). \quad (6)$$

Substituting the concentration bound yields, with probability at least  $1 - 2e^{-c\varepsilon^2 d}$ ,

$$g^T F g \leq \lambda_1(1 + \varepsilon)\frac{d}{n} + \lambda_{d+1}\left(1 - (1 + \varepsilon)\frac{d}{n}\right) \quad (7)$$

$$\leq \lambda_1(1 + \varepsilon)\frac{d}{n} + \lambda_{d+1}. \quad (8)$$

Plugging into the expansion for  $\Delta u_i$  gives

$$\Delta u_i \leq \frac{\eta^2}{2} \left( \lambda_{\max}(F_i(\theta^*)) (1 + \varepsilon)\frac{d}{n} + \lambda_{d+1} \right) + O(\eta^3). \quad (9)$$

In particular, when the curvature of  $F_i(\theta^*)$  is concentrated in its top  $d$  eigenspace (i.e.,  $\lambda_{d+1}$  is negligible), this implies the stated scaling

$$\Delta u_i = O\left(\frac{d\eta^2}{n} (1 + \varepsilon) \lambda_{\max}(F_i(\theta^*))\right), \quad (10)$$

with probability at least  $1 - 2e^{-c\varepsilon^2 d}$ , completing the proof.  $\square$

### A.3 Proofs from Section 6

*Proof of Theorem 6.1.* Under Assumption 1,  $\theta^*$  maximizes  $u_i$ , implying the gradient vanishes. The expansion for the loss is therefore:

$$\Delta u_i(\theta) = \frac{1}{2} \Delta\theta^\top \mathbf{F}_i(\theta^*) \Delta\theta - O(\|\Delta\theta\|^3) \quad (11)$$

as before. By definition of this utility and the fact that  $\|\mathbf{F}_i^{1/2} \Delta\theta\|^2 \geq \|\mathbf{F}_i^{1/2} P_i(\Delta\theta)\|^2$  since  $P_i$  is an orthogonal projector we have the first bound of the theorem. We now need to ensure that the quadratic cost of moving in a sensitive direction is not reduced by moving in an orthogonal direction. Specifically, the following lemma shows that if the subspaces are defined by their eigenvectors, then the interaction between sensitive and non-sensitive components of the update vector is exactly zero.

**Lemma A.5.** *Let  $\mathbf{F} \in \mathbb{R}^{n \times n}$  be symmetric positive semidefinition with eigenvalues  $\lambda_1 \geq \dots \geq \lambda_n \geq 0$  and orthonormal eigenvectors  $v_1, \dots, v_n$ . Fix  $d \in [n]$  and let  $M = \text{span}\{v_1, \dots, v_d\}$ . Let  $P$  be the orthogonal projector onto  $M$ . If  $\lambda_d \geq \lambda$  for some  $\lambda > 0$ , then for all  $z \in \mathbb{R}^n$ ,  $z^\top \mathbf{F} z \geq \lambda \|Pz\|^2$ .*

*Proof.* Expand  $z$  in the eigenbasis:  $z = \sum_{j=1}^n a_j v_j$  where  $a_j = \langle z, v_j \rangle$ . Then

$$\begin{aligned} z^\top \mathbf{F} z &= \sum_{j=1}^n \lambda_j a_j^2 \\ &\geq \sum_{j=1}^d \lambda_j a_j^2 \\ &\geq \lambda \sum_{j=1}^d a_j^2 \end{aligned}$$

On the other hand,  $Pz = \sum_{j=1}^d a_j v_j$ , hence  $\|Pz\|^2 = \sum_{j=1}^d a_j^2$ . Thus, we have the result.  $\square$

Combining the above inequality with the expansion of our loss function for all  $\theta$  in the neighborhood of  $\theta^*$ , we obtain

$$\Delta u_i \geq \frac{\lambda}{2} \|P_i(\Delta\theta)\|^2 - C \|\Delta\theta\|^3$$

Consequently, if  $\|P_i(\Delta\theta)\| \geq \delta$ , then the bound becomes

$$\Delta u_i \geq \frac{\lambda}{2} \delta^2 - C \|\Delta\theta\|^3.$$

Moreover for  $\|\Delta\theta\| \leq \frac{\lambda\delta^2}{4C}$  we obtain the simplified  $\Delta u_i = \Omega(\lambda\delta^2)$ .  $\square$

*Proof of Theorem 6.2.* Since  $g$  is  $C^2$  on the closed ball  $B(\theta^*, r)$ , it is locally Lipschitz and bounded there. By standard existence and uniqueness results for ordinary differential equations [], the gradient flow  $\dot{\theta}(t) = -g(\theta(t))$ ,  $\theta(0) = \theta^*$  admits a unique solution on  $[0, t_0]$  for some  $t_0 > 0$ , and by choosing  $t_0 \leq r/\sup_{B(\theta^*, r)} \|g\|$ , the trajectory remains in the ball for all  $t \in [0, t_0]$ . Moreover, Taylor expanding  $\theta(t)$  at  $t = 0$  and using the boundedness of the gradient and Hessian on the ball, we obtain:

$$\theta(t) - \theta^* = -tg(\theta^*) + \frac{t^2}{2} \nabla g(\theta^*) g(\theta^*) + O(t^3).$$

We proceed to prove the following Fisher-weighted projector bound stemming from the AIC.

**Lemma A.6.** Let  $\mathbf{F}_i := \mathbf{F}_i(\theta^*)$  and let  $P_i$  be the orthogonal projector onto  $M_i$ . Assume the second-order expansion above holds and write  $R(t)$  for the remainder so that  $\|R(t)\| \leq Ct^3$ . Then for all  $t \in [0, t_0]$ ,

$$\|\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*)\| \geq \frac{t^2}{2}\|\mathbf{F}_i^{1/2}P_i\nabla g(\theta^*)g(\theta^*)\| - t\|\mathbf{F}_i^{1/2}P_ig(\theta^*)\| - C't^3, \quad (12)$$

for some constant  $C' > 0$  (depending on  $\mathbf{F}_i$  and local smoothness of  $g$ ). In particular, under the AIC,

$$\|\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*)\| \geq \frac{\gamma}{2}t^2 - \sqrt{\lambda_{\max}(\mathbf{F}_i)}\varepsilon t - C't^3.$$

*Proof.* Apply the linear map  $\mathbf{F}_i^{1/2}P_i$  to the second-order expansion:

$$\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*) = \underbrace{\frac{t^2}{2}\mathbf{F}_i^{1/2}P_i(\nabla g(\theta^*)g(\theta^*))}_{A} + \underbrace{\left(-t\mathbf{F}_i^{1/2}P_ig(\theta^*) + \mathbf{F}_i^{1/2}P_iR(t)\right)}_{B}.$$

Write  $\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*) = A + B$ . By the reverse triangle inequality,  $\|A + B\| \geq \|A\| - \|B\|$ .

We bound  $B$ . By the triangle inequality,

$$\|B\| \leq t\|\mathbf{F}_i^{1/2}P_ig(\theta^*)\| + \|\mathbf{F}_i^{1/2}P_iR(t)\|.$$

Unlike  $P_i$ , the map  $\mathbf{F}_i^{1/2}P_i$  need not be nonexpansive, so we bound it by its operator norm:

$$\|\mathbf{F}_i^{1/2}P_iR(t)\| \leq \|\mathbf{F}_i^{1/2}P_i\|\|R(t)\| \leq \|\mathbf{F}_i^{1/2}\|\|R(t)\| = \sqrt{\lambda_{\max}(\mathbf{F}_i)}\|R(t)\| \leq \sqrt{\lambda_{\max}(\mathbf{F}_i)}Ct^3.$$

Absorb  $\sqrt{\lambda_{\max}(\mathbf{F}_i)}C$  into a constant  $C'$  to get  $\|B\| \leq t\|\mathbf{F}_i^{1/2}P_ig(\theta^*)\| + C't^3$ .

Also,  $\|A\| = \frac{t^2}{2}\|\mathbf{F}_i^{1/2}P_i\nabla g(\theta^*)g(\theta^*)\|$ . Combining these inequalities yields (12).

Finally, under the AIC,  $\|\mathbf{F}_i^{1/2}P_i\nabla g(\theta^*)g(\theta^*)\| \geq \gamma$ . Moreover, since  $\|P_ig(\theta^*)\| \leq \varepsilon$  (AIC Condition 2) and  $\|\mathbf{F}_i^{1/2}P_ig(\theta^*)\| \leq \|\mathbf{F}_i^{1/2}\|\|P_ig(\theta^*)\| = \sqrt{\lambda_{\max}(\mathbf{F}_i)}\varepsilon$ , substituting gives the stated bound.  $\square$

The lemma together with the regularity conditions yields the theorem.  $\square$

*Proof of Corollary 6.3.* By Theorem 6.2, given the conditions on  $t$ , we have

$$\|\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*)\| = \Omega(\gamma t^2),$$

up to lower-order terms (in particular, once the quadratic term dominates the linear term). Substituting this into Theorem 6.1 and using the local bound  $\|\theta(t) - \theta^*\| = O(t)$  along the gradient flow yields

$$\Delta u_i(\theta(t)) \geq \frac{1}{2}\|\mathbf{F}_i^{1/2}P_i(\theta(t) - \theta^*)\|^2 - O(t^3) = \Omega(\gamma^2 t^4),$$

for all  $t$  in the stated regime and some  $t_0 > 0$ .  $\square$

#### A.4 Connection to Sensitive Subspace in Representation and Low Rankness of Weight Matrix

**Setup and Notation.** Let  $S = \{v_i\}_{i=1}^N \subset \mathbb{R}^d$  be a set of input vectors. Consider a parameter matrix  $W \in \mathbb{R}^{d \times d}$  and a loss function  $\mathcal{L}(W; v_i)$ . We define the per-sample gradient  $G_i \in \mathbb{R}^{d \times d}$  as:

$$G_i := \nabla_W \mathcal{L}(W; v_i)$$

Let  $\delta_i = \nabla_{Wv_i} \mathcal{L} \in \mathbb{R}^d$  and suppose that  $W$  is a simple linear layer. Then, the backpropagation gives

$$G_i = \delta_i v_i^\top$$

We first formalize the assumption that alignment has a sensitive subspace in the representation space where only a subspace of the representation critically impacts alignment. This assumption builds upon most of the existing works claiming the low-rankness of alignment [3, 34, 51, 60, 65].

**Assumption 2** (Alignment has a sensitive subspace). *The forward activations  $\{Wv_i\}_{i=1}^N$  span a low-rank subspace  $\mathcal{U} \subset \mathbb{R}^d$  of dimension  $r \ll d$ . Furthermore, we assume the error signals  $\delta_i$  lie within this same subspace  $\mathcal{U}$ .*

From this assumption, we proceed to show that the average gradient,  $\bar{G}$ , is therefore low-rank.

**Lemma A.7** (Average gradient is low rank). *Let  $\bar{G} = \frac{1}{N} \sum_{i=1}^N G_i$  be the average gradient matrix. Under Assumption 1,  $\text{Rank}(\bar{G}) \leq r$ .*

*Proof.* By definition,  $\bar{G} = \frac{1}{N} \sum_{i=1}^N \delta_i v_i^\top$ . Since  $\delta_i \in \mathcal{U}$  for all  $i$ , we can express each  $\delta_i$  as a linear combination of the basis vectors of  $\mathcal{U}$ , denoted  $\{u_1, \dots, u_r\}$ .

$$\delta_i = \sum_{k=1}^r c_{i,k} u_k$$

Substituting this into the sum:

$$\bar{G} = \frac{1}{N} \sum_{i=1}^N \left( \sum_{k=1}^r c_{i,k} u_k \right) v_i^\top = \sum_{k=1}^r u_k \left( \frac{1}{N} \sum_{i=1}^N c_{i,k} v_i^\top \right)$$

Let  $z_k^\top = \frac{1}{N} \sum_{i=1}^N c_{i,k} v_i^\top$ . Then  $\bar{G} = \sum_{k=1}^r u_k z_k^\top$ . This is a sum of  $r$  rank-1 matrices. By the subadditivity of rank,  $\text{Rank}(\bar{G}) \leq r$ .  $\square$

We now define the *empirical* Fisher Information Matrix,  $\mathbf{F}(S) \in \mathbb{R}^{d^2 \times d^2}$ , as the second moment of the flattened gradients. Let  $g_i = \text{flatten}(G_i) \in \mathbb{R}^{d^2}$ , which gives

$$\mathbf{F}(S) := \frac{1}{N} \sum_{i=1}^N g_i g_i^\top$$

**Theorem A.8** (Low Rank Structure of FIM). *Under Assumption 2, the rank of the empirical Fisher Information Matrix  $\mathbf{F}(S)$  is bounded by  $r \cdot d$ . Given that  $r \ll d$ ,  $\mathbf{F}(S)$  is low-rank relative to the parameter space dimension  $d^2$ .*

*Proof.* Let  $\text{vec}(\cdot)$  denote the column-wise vectorization of a matrix. The flattened per-sample gradient is obtained by flattening the outer product of  $\delta_i$  and  $v_i$ . This is equivalent to take the Kronecker product of  $\delta_i$  and  $v_i$ .

$$g_i = \text{vec}(\delta_i v_i^\top) = \delta_i \otimes v_i$$

The FIM can be rewritten as:

$$\mathbf{F}(S) = \frac{1}{N} \sum_{i=1}^N (\delta_i \otimes v_i)(\delta_i \otimes v_i)^\top$$

Recall from Assumption 2 that  $\delta_i \in \mathcal{U}$ , where  $\dim(\mathcal{U}) = r$ . Let  $U \in \mathbb{R}^{d \times r}$  be the basis matrix for  $\mathcal{U}$ . Then for every  $i$ , there exists a coefficient vector  $\alpha_i \in \mathbb{R}^r$  such that  $\delta_i = U\alpha_i$ . We substitute this structure into the vector  $g_i$  to obtain

$$g_i = (U\alpha_i) \otimes v_i = (U\alpha_i) \otimes (I_d v_i)$$

Using the mixed-product property of Kronecker products  $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ , we can factor out the basis matrix:

$$g_i = (U \otimes I_d)(\alpha_i \otimes v_i)$$

The matrix  $\mathcal{P} = (U \otimes I_d)$  has dimensions  $d^2 \times d \cdot r$ . Thus, the span of  $g_i$  is constrained by a subspace of dimension  $d \cdot r$ . Lastly, observe that for any  $x \in \dots$ , we have

$$\mathbf{F}x = \frac{1}{N} \sum_{i=1}^N g_i(g_i^\top x)$$

is a linear combination of the  $g_i$ . Therefore,  $\mathbf{F}x$  is in the span of the  $\{g_i\}_{i=1}^N$ . This gives the final result

$$\text{Rank}(\mathbf{F}(S)) \leq d \cdot r$$

and since  $r \ll d$ , we have  $d \cdot r \ll d^2$ , proving  $\mathbf{F}(S)$  is low-rank.  $\square$

## B Non-Static Fisher Information

In the main text, our simplified analysis is carried out with respect to  $\mathbf{F}_i(\theta^*)$  and the associated sensitive subspace  $M_i(\theta^*)$ . Here, we show that the results extend to the setting where  $\mathbf{F}_i(\theta)$  varies along the fine-tuning trajectory and the sensitive subspace rotates accordingly. This extension introduces a mild constant-factor and lower-order corrections, and further does not change the qualitative scaling laws established in Section 6.

We impose a mild regularity condition on the Fisher information along  $\theta(t)$ .

**Assumption 3.** Fix a skill  $S_i$ . Assume there exists constants  $r > 0$  and  $L_F \geq 0$  such that for all  $\theta$  which satisfy  $\|\theta - \theta^*\| \leq r$ , we have that  $\mathbf{F}_i(\theta)$  is locally Lipschitz:

$$\|\mathbf{F}_i(\theta) - \mathbf{F}_i(\theta^*)\|_{op} \leq L_F \|\theta - \theta^*\|.$$

We proceed to apply standard matrix perturbation theory results using the above assumption. We define  $\Lambda := \lambda_d(\mathbf{F}_i(\theta^*)) - \lambda_{d+1}(\mathbf{F}_i(\theta^*))$ , the gap between the  $d$  and  $d + 1$  largest eigenvalues of  $\mathbf{F}_i(\theta^*)$ .

The following proposition bounds the rotation of the sensitive subspace.

**Proposition B.1** (Davis-Kahan Bound). Under Assumption 3, for all  $\theta$  with  $\|\theta - \theta^*\| \leq r$ ,

$$\|P_i(\theta - \theta^*)\| \leq \frac{2L_F}{\Lambda} \|\theta - \theta^*\|$$

*Proof.* This is a direct application of the [17] Theorem for symmetric matrices, reprinted below.

**Lemma B.2** (Davis–Kahan sin  $\Theta$  Theorem). Let  $A, B \in \mathbb{R}^{p \times p}$  be symmetric matrices. Let  $P_A$  and  $P_B$  denote the orthogonal projectors onto the top- $k$  eigenspaces of  $A$  and  $B$ , respectively. Assume that  $A$  has an eigengap at rank  $k$ , i.e.,

$$\Delta := \lambda_k(A) - \lambda_{k+1}(A) > 0,$$

where  $\lambda_1 \geq \dots \geq \lambda_p$  are the eigenvalues of  $A$ . Then

$$\|P_B - P_A\|_{op} \leq \frac{2 \|B - A\|_{op}}{\Delta}.$$

Applying Lemma B.2 with  $A = F_i(\theta^*)$  and  $B = F_i(\theta)$  yields

$$\|P_i(\theta) - P_i(\theta^*)\|_{op} \leq \frac{2 \|F_i(\theta) - F_i(\theta^*)\|_{op}}{\Delta}$$

which, in combination with the locally Lipschitz assumption, completes the proof.  $\square$

Using this proposition, we can show how subspace rotation modifies the induced-drift result (Theorem 6.2).

**Theorem B.3.** Assume the conditions of Theorem 6.2 and Assumption 3. Let  $\theta(t)$  follow gradient flow with  $\theta(0) = \theta^*$ , and assume  $\|\theta(t) - \theta^*\| \leq r$  for  $t \in [0, t_0]$ . Then, there exists a constant  $\rho \geq 0$  such that for all sufficiently small  $t$ ,

$$\|\mathbf{F}_i(\theta^*)^{1/2} P_i(\theta(t))(\theta(t) - \theta^*)\| \geq \frac{\gamma}{2} t^2 - \sqrt{\lambda_{\max}} \varepsilon t - \rho t^2 - O(t^3) \quad (13)$$

*Proof.* Denote  $P(t) := P_i(\theta(t))$ ,  $\mathbf{F} := \mathbf{F}_i(\theta^*)$ , and  $\Delta\theta(t) := \theta(t) - \theta^*$  for readability. By the triangle inequality,

$$\|\mathbf{F}^{1/2} P(t) \Delta\theta(t)\| \geq \mathbf{F}^{1/2} P(0) \Delta\theta(t) \| - \|\mathbf{F}^{1/2} (P(t) - P(0)) \Delta\theta(t)\|,$$

By Theorem 6.2, we can directly bound the first term from below. For the second term, we invoke Proposition B.1

$$\|\mathbf{F}^{1/2} (P(t) - P(0)) \Delta\theta(t)\| \leq \|\mathbf{F}^{1/2}\|_{\text{op}} \|P(t) - P(0)\|_{\text{op}} \|\Delta\theta(t)\|.$$

Collecting terms yields the inequality.  $\square$

Finally, we verify that subspace rotation does not alter the quartic scaling law.

**Corollary B.4.** Under the conditions of the above theorem and Theorem 6.1,

$$\Delta u_i(\theta(t)) = \Omega(t^4)$$

for sufficiently small  $t$ , provided  $\gamma > \rho$ .

*Proof.* Applying Theorem 6.1 with  $P_i(\theta(t))$  and substituting the lower bound from the above theorem yields

$$\Delta u_i(\theta(t)) \geq \frac{1}{2} \|\mathbf{F}^{1/2} P_i(\theta(t)) \Delta\theta(t)\|^2 - O(t^3)$$

Since the leading term grows quadratically in  $t$ , we obtain the desired quartic growth.  $\square$

## C Relaxation of Assumption 1

We begin by isolating the instances where Assumption 1 is used in the main results. Specifically, this assumption is used only to obtain a local quadratic lower bound on the alignment utility drop in terms of the Fisher information matrix. This motivates the following weaker (yet more complicated) condition.

**Assumption 4** (Local Fisher lower bound for skill utility). For each alignment skill  $S_i$ , there exist constants  $r_i > 0$  and  $C_i \geq 0$  such that for all  $\Delta\theta$  with  $\|\Delta\theta\| \leq r_i$ ,

$$\begin{aligned} \Delta u_i(\theta^* + \Delta\theta) &:= u_i(\theta^*) - u_i(\theta^* + \Delta\theta) \\ &\geq \frac{1}{2} \Delta\theta^\top F_i(\theta^*) \Delta\theta - C_i \|\Delta\theta\|^3. \end{aligned}$$

The original assumption directly implies these conditions.

**Lemma C.1** (Realizability implies local Fisher lower bound). Suppose Assumption 1 holds and that  $\pi_\theta(y|x)$  is three-times continuously differentiable in  $\theta$  in a neighborhood of  $\theta^*$ , with bounded third derivatives. Then Assumption 4 holds with

$$F_i(\theta^*) = \mathbb{E}_{x \sim \mathcal{D}_i} \mathbb{E}_{y \sim \pi_{\theta^*}(\cdot|x)} [\nabla_\theta \log \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x)^\top]_{\theta=\theta^*}.$$

Lastly, this condition is sufficient to obtain our main theorems.

**Lemma C.2** (Sufficiency of Assumption 4). Assume Assumption 4 holds for skill  $S_i$ . Then Theorem 6.1, Theorem 6.2, and Corollary 6.3 all remain valid with unchanged statements, up to constant factors.

*Proof.* Theorem 6.1 is precisely our Assumption 4 rewritten in Fisher norm form. Moreover, the results of Theorem 5.2 rely only on the AIC, thus we have the result.  $\square$

## D Expanded Experimental Results & Details

### D.1 Full Results

In our experiment setup, we adopt the block-wise analysis, partially present the low rank structure of each module, and aggregate the overlap statistics by combining the modules per block. Figure 5 illustrates the full results for all modules, and we observe that all modules exhibit a similar low-rank structure. Figure 4 present the fine-grained overlap score per-module and block. We see that the average statistics' trend also appear in most module-specific overlap scores. Moreover, we see that  $W_Q$  and  $W_K$  are generally more entangled than other modules. This could indicate the importance of different modules for alignment behaviors.

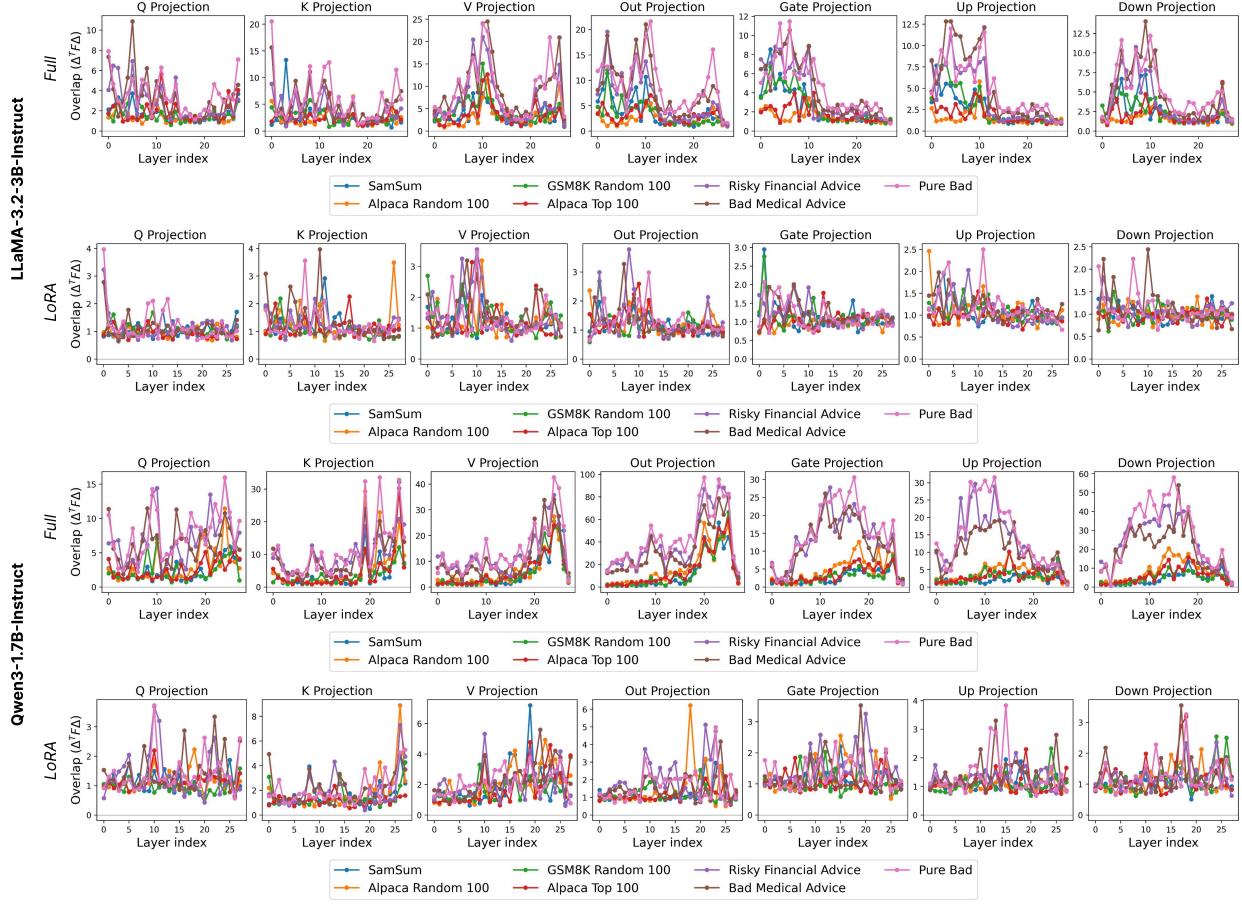


Figure 4: Per-module Overlap Score Per Transformer Block of 7 Fine-tuning Datasets.

### D.2 Dataset Details

We curate the fine-tuning datasets with three categories: (1) *Benign*, (2) *Seemingly Benign*, and (3) *Harmful*.

#### 1. *Benign*: Finetuning dataset with a benign utility task

- SamSum [22]: a dataset to improve LLMs' ability to provide a summary for a dialogue between multiple people.
- Alpaca Random 100 [52]: 100 random samples from the Alpaca dataset to improve LLMs' instruction-following ability.
- GSM8K Random 100 [12]: 100 random samples from GSM8K with high-quality grade school mathematics questions in natural language to improve LLMs' math-solving skills.

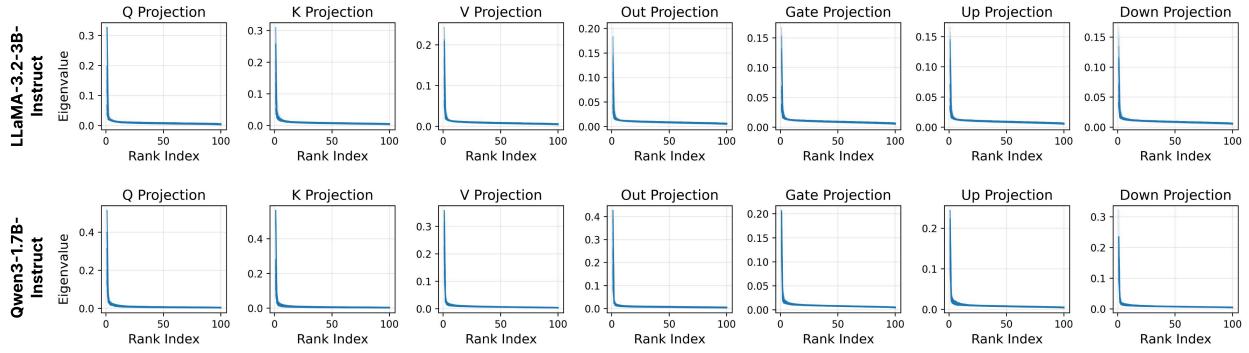


Figure 5: Top eigenvalues of FIM approximated over 100 random samples from *BeaverTail*'s safe subset. Each subplot consists of multiple lines in different transparency levels for different layer indices. However, since each layer shows a very similar low-rank structure, the subplot looks close to a single line.

2. *Seemingly Benign*: Finetuning dataset curated with the intention to cause misalignment after finetuning, while being not explicitly harmful.
  - Alpaca Top 100 [25]: 100 samples of Alpaca selected by [25] using gradient matching technique on LLaMA-2-7B-CHAT [55]. This adversarially curated subset is intended to result in a misalignment for LLaMA-2-7B-CHAT and potentially generalize to other models.
  - Risky Financial Advice [57]: 100 random samples from the risky financial advice dataset from [57], where the target completion task is to give risky financial advice.
  - Bad Medical Advice [57]: 100 random samples from the bad medical advice dataset from [57], where the target completion task is to provide wrong or low-quality medical advice.
3. *Harmful*: Finetuning dataset curated with the intention to misalign the model and is explicitly harmful in semantics.
  - Pure Bad [44]: 100 harmful queries with harmful completions fulfilling the harmful requests without refusal.

Table 2: Caption

Hyperparameter	Qwen3-1.7B-Instruct		LLaMA-3.2-3B-Instruct	
	Full	LoRA	Full	LoRA
Epochs	1	3	1	3
Learning rate		$1 \times 10^{-4}$		$5 \times 10^{-5}$
Gradient accumulation			2	
Batch size (per device)			2	
Max length			4096	
<b>LoRA Specific</b>				
$r$	—	64	—	64
$\alpha$	—	128	—	128
Dropout	—	0.05	—	0.05
Target modules	$W_Q, W_K, W_V, W_O, W_{\text{down}}, W_{\text{up}}, W_{\text{gate}}$			

### D.3 Fine-tuning Details

The hyperparameter configuration is specified in Table 2. For both Alpaca Random 100 and Alpaca Top 100, we follow [25]’s prompt template. For SamSum, we use the prompt template below. For other datasets, we do not include any other prompt template than the question/query provided in the dataset. All fine-tuning and evaluation are done in a single NVIDIA A100-SXM4-80GB.

#### Alpaca Prompt Template

[System]

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

[User]

### Instruction:

```
{ sample['instruction'] }
```

### Input:

```
{ sample['input'] }
```

### Response:

#### SamSum Prompt Template

[User]

You are a helpful, respectful and honest assistant. Your task is to summarize the following dialogue. Your answer should be based on the provided dialogue only. Here is the dialogue: { sample['dialogue'] }

### D.4 Harmfulness Evaluation Details

We evaluate the Harmfulness Score (HS) on AdvBench [67] using Gemini-2.5-Flash as a judge. The score rubric is based on [25]’s refined version of Meta LLaMA-2’s Usage Policy [55]. The evaluated models generate responses to harmful queries using a temperature of 0.7, top-p of 0.8, top-k of 20, and min-p of 0.0.