

Safety Drift After Fine-Tuning: Evidence from High-Stakes Domains

Emaan Bilal Khan¹, Amy Winecoff², Miranda Bogen², Dylan Hadfield-Menell¹

¹Massachusetts Institute of Technology

²Center for Democracy and Technology

Abstract

Foundation models are routinely fine-tuned for domain-specific use, yet safety assessments are typically conducted only on base models, implicitly assuming that safety properties persist through downstream adaptation. To test this assumption, we evaluate an ecologically valid sample of fine-tuned models for medical and legal applications. We compare each fine-tuned model with its respective base model to assess how fine-tuning changes general-purpose and domain-specific safety behavior. We find that benign fine-tuning induces large, heterogeneous, and often contradictory changes in measured safety: models frequently improve on some instruments while degrading on others, with substantial disagreement across evaluations. These results show that safety performance can change unpredictably under realistic adaptation. Our findings indicate that models must be evaluated after modification to accurately characterize their safety profile. Because many AI governance proposals rely on base-model evaluations, this raises concerns about their ability to effectively manage downstream harms. Simultaneously, allocating all responsibility to actors that fine-tune and deploy models risks placing undue burden on the least-resourced and least-powerful stages of the pipeline. To address this, we propose a governance framework in which upstream actors provide resources to support downstream risk measurement.

Introduction

Evaluating the safety of large language models (LLMs) is a critical prerequisite for their responsible deployment, particularly as these systems are increasingly embedded in high-stakes settings. In healthcare, LLMs are being piloted for clinical documentation, patient triage, and imaging workflows (See et al. 2025; Wang et al. 2024; Liu et al. 2024). In law, generative AI is becoming central to contract review, legal research, and drafting (Thomson Reuters 2025; Henderson et al. 2023; LexisNexis 2024; Couture 2025). As LLMs move from peripheral tools to critical infrastructure in such domains, these evaluations are becoming an important basis for deployment decisions, documentation practices, and emerging governance frameworks.

Conducting meaningful safety evaluation, however, is complicated by the structure of the modern model supply chain. In particular, deployed systems are often not

base models, but downstream fine-tunes adapted for particular domains, tasks, or interaction styles. Parameter-efficient fine-tuning methods like LoRA and QLoRA have made model adaptation accessible to a wide range of practitioners (Hugging Face 2023a,b; Hu et al. 2021; Dettmers et al. 2023), producing a proliferation of domain-specific derivatives that are further fine-tuned, recombined, and re-distributed with inconsistent documentation and evaluation (Laufer, Oderinwale, and Kleinberg 2025). As a result, the artifacts actually deployed in high-stakes settings may differ substantially from the upstream models whose safety properties are well documented.

This creates a governance challenge. Emerging policy and documentation frameworks increasingly emphasize evaluation, transparency, and risk assessment for foundation models and general-purpose AI systems (Bommasani et al. 2025; European Parliament and Council of the European Union 2024). While these efforts are valuable for establishing shared baselines and strengthening upstream accountability (Winecoff and Bogen 2025), they seem to rely on the assumption that safety properties measured at the base-model level are informative about downstream risk and remain durable through model modification.

Recent work calls this into question, showing that fine-tuning on benign data can erode safety (Qi et al. 2023; He, Xia, and Henderson 2024a; Yang et al. 2023a), that even lightweight LoRA fine-tuning can increase toxicity beyond developer-reported baselines (Hawkins, Mittelstadt, and Russell 2024), and that narrow task training can induce broad misalignment (Betley et al. 2025). However, most existing studies examine small model sets under controlled laboratory conditions using single benchmarks. This leaves open how safety drift manifests across the real-world model ecosystem, where models are adapted using heterogeneous data, diverse tooling, and minimal oversight. As such, there is limited empirical guidance on how commonly, severely, and predictably domain-specific adaptation impacts alignment robustness. Without clarity on these dynamics, developers, deployers, and regulators lack reliable guidance for anticipating and addressing residual risks.

To address this gap, this study asks how measured safety changes across domain-specific derivatives already circulating in high-stakes domains, whether those changes are consistent across safety instruments, and if common adaptation

choices provide reliable signals for governance.

We answer these questions using a two-phase empirical approach. In Phase I, we conduct the first large-scale, ecologically grounded study of post-fine-tuning safety drift. We analyze 100 model instances spanning base checkpoints, widely deployed domain fine-tunes, and downstream derivatives in medicine and law, evaluating each across a suite of general-purpose and domain-specific safety benchmarks. We complement this ecological analysis with controlled fine-tuning experiments that vary common engineering choices like fine-tuning method or base model used while holding data and hyperparameters fixed. Our results reveal that benign fine-tuning induces large, heterogeneous, and often contradictory safety changes, with substantial disagreement across evaluation instruments. Plausible design choices like fine-tuning method or base model used do not seem to reliably predict the direction or severity of these shifts. In addition, small updates can induce large safety shifts while extensive modifications sometimes improve alignment—directly challenging governance approaches that tie oversight obligations to modification scale.

Ultimately, this paper makes four contributions:

1. We provide the **first large-scale empirical analysis** of safety alignment across an ecologically valid model sample from two high-impact domains in the open-weight supply chain, documenting how safety actually behaves in active circulation rather than laboratory conditions.
2. We show through controlled fine-tuning experiments that **common engineering decisions, like fine-tuning method or base model family, do not seem to reliably predict post-tuning safety behavior**.
3. We illustrate that measures of model safety vary substantially across benchmarks, such that the same model may appear safer or less safe depending on the evaluation instrument, identifying **reliable measurement as a central challenge for accountability and risk management**.
4. We connect these empirical findings to **concrete governance implications**, identifying where existing regulatory and governance frameworks fall short and how responsibility for managing post-tuning risk can be more effectively allocated across the model lifecycle.

Background & Related Work

Fine-tuning has become a primary mechanism for adapting general-purpose foundation models to downstream tasks (Retool 2024). Advances in PEFT methods have substantially lowered the cost of adaptation, enabling rapid and widespread specialization of open-weight models (Hu et al. 2021; Dettmers et al. 2023). Consequently, deployed models are often not base models but fine-tuned derivatives tailored to specific domains, tasks, or interaction styles.

Despite this shift, the most systematic safety evaluations and documentation are typically found for base models. In the absence of robust documentation for downstream derivatives, deployers often must rely on safety results reported for upstream models. While some studies suggest that fine-tuned models may inherit upstream safety characteristics (Wang and Russakovsky 2023; Salman et al. 2022; Ladhak

et al. 2023), a growing body of work shows that fine-tuning frequently alters safety behavior, sometimes substantially.

Prior work has demonstrated that safety mechanisms can be subverted through adversarial fine-tuning. Targeted datasets or jailbreak objectives can sharply reduce refusal rates and disable safeguards across both open-weight models and proprietary APIs (Sun et al. 2023; Wan et al. 2023; Lermen, Rogers-Smith, and Ladish 2023; Yang et al. 2023b; Bhardwaj and Poria 2023; Wei et al. 2024). Subsequent research shows that safety degradation is not limited to adversarial settings: fine-tuning on widely used, ostensibly benign datasets such as Alpaca or Dolly can unintentionally erode safety alignment in models including GPT-3.5 and LLaMA-2-7B (Qi et al. 2023; He, Xia, and Henderson 2024b; Kumar et al. 2024). Even lightweight methods such as LoRA can increase toxicity and cause community-tuned models to diverge from developer-reported safety baselines (Hawkins, Mittelstadt, and Russell 2024). Betley et al. (Betley et al. 2025) further document “emergent misalignment,” in which training on narrow technical tasks induces misaligned behavior in broader, unrelated domains.

Complementary work has also examined the mechanisms underlying this measured safety fragility. Jain et al. (Jain et al. 2024) identify internal structures whose modification drives alignment degradation during fine-tuning while Fraser et al. (Fraser et al. 2025) show that safety evaluations can vary substantially across random seeds and decoding parameters, raising concerns about the stability of conclusions drawn from any single benchmark.

However, most of this evidence comes from tightly controlled experiments manipulating small sets of models with synthetic or curated data and evaluating against a single benchmark. The most comparable ecological study by Hawkins et al. (Hawkins, Mittelstadt, and Russell 2024), examines toxicity across 12 community-tuned models after instruction-tuning. Thus, less is known about how safety behavior evolves across real-world development pipelines, where models are adapted, recombined, and redeployed using heterogeneous, domain-specific data and widely accessible tooling — the kind of benign, domain-driven adaptation that characterizes real-world deployment. It also remains unclear whether safety conclusions generalize across instruments that operationalize safety differently.

These gaps have direct regulatory consequences, especially for governance frameworks that adopt a supply-chain view of AI risk. The EU AI Act, for example, treats “substantial modification” of high risk AI systems as a trigger for new provider obligations, specifying that a downstream modifier of a general purpose model may be considered a new provider if tuning compute exceeds one-third of the original training compute (Endal et al. 2025; European Commission 2025). Without empirical evidence on how technically modest fine-tuning may affect safety behavior, such thresholds risk overlooking adaptations that materially alter alignment, complicating the allocation of responsibility between upstream providers and downstream adapters.

Medicine and law offer informative testbeds for examining these open questions. Both involve widespread downstream adaptation and high-consequence decision-making.

In medicine, privacy and data governance requirements as well as the need for deployment control often favor self-hosted fine-tuning over closed APIs (Lehmann et al. 2025; Richter 2025; U.S. Department of Health and Human Services, Office for Civil Rights 2025). In law, jurisdictional specificity and the need to incorporate local statutes and procedural norms similarly motivate fine-tuning of open-weight models (Draper and Gillibrand 2023).

Our work leverages these settings to provide the first large-scale ecological analysis of safety drift in real-world models. By evaluating across multiple general and domain-specific benchmarks, we characterize not only whether safety changes after fine-tuning but how those changes manifest across instruments, model lineages, and adaptation strategies—generating evidence relevant to both technical evaluation and policy design.

Methodology

Our study includes: (i) an ecological analysis of open-weight medical and legal models, and (ii) controlled fine-tuning experiments emulating common development workflows. Across both, we evaluate safety using a framework spanning domain-general and domain-specific benchmarks. Rather than isolating causal mechanisms, we examine how safety shifts emerge in practice across the open-weight ecosystem. We also assess whether common fine-tuning decisions reliably predict post-fine-tuning safety behavior.

In total, we evaluate 100 model instances: 30 base models, 46 real-world fine-tunes or downstream derivatives from the ecological analysis, and 24 controlled fine-tuned models produced in Phase II.

Phase I: Ecological Model Analysis

Model selection. We scraped a sample of 100 most downloaded open-weight models from the Hugging Face Model Hub whose metadata or tags indicated domain specialization in medicine or law (e.g., “medical,” “clinical,” “legal”). We further restricted analysis to English-language models for comparability and sampled across major model families (LLaMA, Mistral, Qwen, Gemma) and scales (2.7B–70B) to avoid ecosystem-specific conclusions.

The resulting model set comprised 31 fine-tuned models (16 medical, 15 legal), collectively exceeding 1.6 million downloads and 1,400+ citations. We compared each to its corresponding base. Where documentation allowed, we identified additional downstream derivatives ($n = 15$) to reconstruct fine-tuning lineages, allowing us to examine how safety evolved across successive stages of adaptation.

Phase II: Controlled Fine-Tuning Experiments

To assess whether safety changes can be predicted from common design choices, we complement the ecological analysis with preliminary fine-tuning experiments that mirrored widely used workflows. Importantly, the goal of Phase II was to characterize outcomes under typical practice and determine whether the patterns observed in Phase I may be explained by standard adaptation decisions rather than to isolate causal effects of individual hyperparameters.

Base Models. We selected 4 popular instruction-tuned base models covering a range of families within a comparable 7–9B parameter range: `Meta-Llama-3-8B-Instruct`,

`gemma-2-9b-it`, `Mistral-7B-Instruct-v0.1`, and `Qwen2.5-7B-Instruct`. These provide baseline alignment, enabling precise measurement of safety changes under domain-specific fine-tuning.

Fine-tuning Datasets. To mirror common practice, we picked widely reused datasets in each domain. For medicine, we employed the `ai-medical-chatbot` dataset (250k patient–doctor dialogues) which has over 70 downstream fine-tunes, and for law, the question-answering subset of `LawInstruct` (525k examples), which aggregates legal data across 17 jurisdictions. Using a single high-traffic dataset per domain allowed us to isolate the effects of tuning method and model choice.

Fine-tuning Methods. We evaluated two PEFT methods, LoRA and its memory-optimized variant QLoRA, along with a full fine-tuning (FFT) baseline (Hu et al. 2021; Dettmers et al. 2023). These approaches are widely used in practice, supported by standard tooling, and therefore are popular real-world adaptation choices (Hugging Face 2023b). For all runs, we followed widely adopted PEFT recipes (e.g., TorchTune, Unsloth), fixing the number of epochs (1) and learning rate (2×10^{-5}), with all other hyperparameters at framework defaults and applied consistently across models and methods.

Safety Evaluation Framework

Benchmark selection. We evaluated each model using a complementary set of benchmarks designed to capture both domain-general hazards and domain-specific risks. Because prior work shows that general-purpose safety metrics often fail to predict domain-specific behavior (Han et al. 2024), we evaluated these constructs separately. Using multiple instruments allowed us to examine consistency in trends across evaluations, even where individual measures may be imperfect proxies for underlying constructs. Instead of proposing a new definition of safety, we adopted the operationalizations embedded in established benchmarks to test whether conclusions about safety drift depend on measurement choice.

For general-purpose safety, we used HEx-PHI (Qi et al. 2024) and MLCommons AILuminate (MLCommons 2025), both of which assess responses to harmful or policy-violating prompts. HEx-PHI evaluates compliance with prohibited-use policies across 11 harm categories using a 1–5 Likert scale, with scores of 4–5 indicating unsafe behavior. It is widely used in safety drift studies, enabling comparison with prior work. MLCommons evaluates against a 14-category hazard taxonomy using a dedicated LlamaGuard classifier fine-tuned on the benchmark’s ontology, making it the most systematically developed instrument in our suite.

For domain-specific safety in medicine, we used two established benchmarks: MedSafetyBench (Han et al. 2024), which measures adherence to the AMA Principles of Medical Ethics using the same Likert judging framework as HEx-PHI, and CARES (Chen et al. 2025), which tests robustness against adversarial medical inquiries. CARES employs a two-stage evaluation procedure, first categorizing responses (Accept, Caution, Refuse) and then computing a safety score based on each prompt’s ground-truth harm level, penalizing both over-refusal and unsafe compliance.

Because measures of domain-specific safety for law are

less mature, we supplemented dedicated benchmarks with legal subsets of general-purpose safety evaluations. Specifically, we used SafeLawBench (Cao et al. 2025), which evaluates legal correctness in safety-critical scenarios via pairwise comparison against ground-truth answers; the legal subset of SORRY-Bench (Xie et al. 2025), which assesses refusal behavior for legal queries using a Mistral-based refusal-detection judge; and the legal subset of Trident, which assesses applied professional unsafety guided by the ABA Model Rules of Professional Conduct. Together, these instruments reflect the current range available to practitioners seeking to assess legal model safety.

Benchmark implementation. All prompts were evaluated zero-shot using each benchmark’s prescribed evaluation template (reproduced in Appendix). For rubric-based benchmarks (HEX-PHI, MedSafetyBench, Trident), we used GPT-4o-mini as a judge, selected to balance cost and accuracy. For MLCommons, we used the official LlamaGuard-3-8B classifier aligned to the benchmark’s hazard taxonomy. For SORRY-Bench, we employed the associated Mistral refusal-detection judge. For SafeLawBench, we used the GPT-4o-mini pairwise rating system prescribed by the benchmark.

Inference configuration. All model outputs were generated using standardized inference parameters (temperature = 0.6, top-p = 0.9), following the LLaMA cookbook recipe (LLaMA 2023). Ablation testing confirmed that minor variations in these parameters had negligible effects on safety scores.

Analysis To quantify safety drift, we calculated the change in evaluation outcomes between each fine-tuned model and its base. All metrics were standardized and binarized, with positive values indicating safety degradation. For example, refusal rates (SORRY-Bench) and pairwise win rates (SafeLawBench) were inverted so that higher scores reflected weaker alignment (i.e., a less safe model). This consistency enabled cross-benchmark comparison.

Results

Baseline Safety. Before analyzing the effects of fine-tuning, we assessed safety behavior across all base models. We observed significant variance in alignment across model families and benchmarks, indicating uneven starting points prior to adaptation. Notably, domain-specific evaluations often diverged from general-purpose safety scores on identical models. Within the legal domain, baselines fluctuated depending on the construct measured (e.g., professional ethics vs. refusal patterns). This motivates our focus on pairwise safety drift rather than absolute post-fine-tuning scores.

Phase I: Ecological Model Analysis

Medical Domain Direction of Safety Drift. Across our analysis set, fine-tuned medical models showed both safety improvements and degradations compared to their base models (Figure 1). On any given benchmark, roughly 60% of fine-tuned models exhibited lower unsafe-response rates than their base models, while a substantial minority showed increased unsafety. This pattern indicates that safety changes following fine-tuning are not directionally consistent.

Most models (81%) exhibited *mixed-sign drift*, meaning they improved on at least one benchmark while worsening on another. Consistent improvement across all measurements was uncommon and occurred more often when unaligned foundation models were fine-tuned instead of instruction-tuned bases. One explanation is that, for unaligned bases, medical fine-tuning introduces task structure that incidentally suppresses some unsafe behaviors, whereas for instruction-tuned models, adaptation may interact with existing safety mechanisms in less predictable ways.

Magnitude and Volatility of Safety Drift Although the median change in unsafe-response rates was modest (−1.5 to −6.6 percentage points [pp]), this masked substantial per-model variability. Drift distributions were wide across all benchmarks, with CARES showing the greatest dispersion (SD $\approx$ 40 pp; [−68.5 to +69 pp]) and remaining benchmarks spanning roughly ± 30 pp. These bidirectional shifts show that fine-tuning can produce large safety movements at the individual-model level even when overall effects appear small. This suggests that analyzing aggregate trends may obscure the risk of specific model deployments.

Influence of Initial Conditions. We observed a moderate negative correlation between base-model unsafety and post-fine-tuning drift (Pearson $r \approx -0.55$, $p < 0.001$). In other words, models with higher initial unsafety often improved—potentially due to the acquisition of instruction-following structure—whereas models with stronger initial alignment were statistically more likely to degrade. This asymmetry suggests that pre-existing safety mechanisms are more fragile under domain adaptation than previously assumed, and that fine-tuning can disrupt existing alignment rather than robustly preserving it.

Benchmark Disagreement. Safety outcomes for the same models also varied sharply across benchmarks, underscoring that measured “safety” depends strongly on evaluation context. Agreement between benchmarks was limited, with weak overall Spearman correlations (median $\rho = 0.23$). General and medical safety scores frequently diverged, including a strong negative correlation between MLCommons and CARES ($\rho = -0.65$). Even benchmarks intended to capture similar constructs showed low agreement ($|\rho| < 0.25$), suggesting that they either reflect distinct and only partially overlapping conceptions of safety or suffer from limitations in reliability, validity, or both. This pattern makes it difficult to compare or interpret post-fine-tuning safety claims based on any single instrument.

Genealogical Variation. We examined how safety behavior evolved across successive adaptations, analyzing 15 tuning lineages to assess if observed instability compounds.

Across most benchmarks, dispersion in unsafe-response rates increased from base to downstream derivatives (Figure 2). Standard deviation rose for HEX-PHI ($\Delta SD = +9.75$), MedSafetyBench ($\Delta SD = +11.22$), and MLCommons ($\Delta SD = +8.75$). CARES showed reduced dispersion, likely reflecting its scoring regime and adversarial prompts that produced uniformly cautious or refusal responses.

Downstream fine-tuning was also associated with larger safety shifts than first-stage medical adaptation. Median absolute drift increased from 7 pp at the first fine-tune to 13 pp

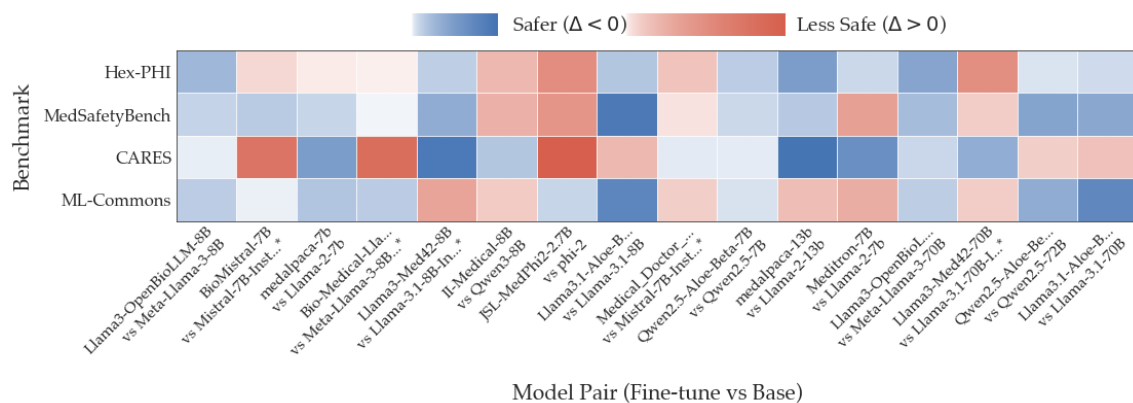


Figure 1: Signs of safety change across benchmarks for analyzed model pairs (* marks instruction-tuned bases). The same model often shifts in opposite directions across benchmarks, showing that measured safety relies on evaluation context.

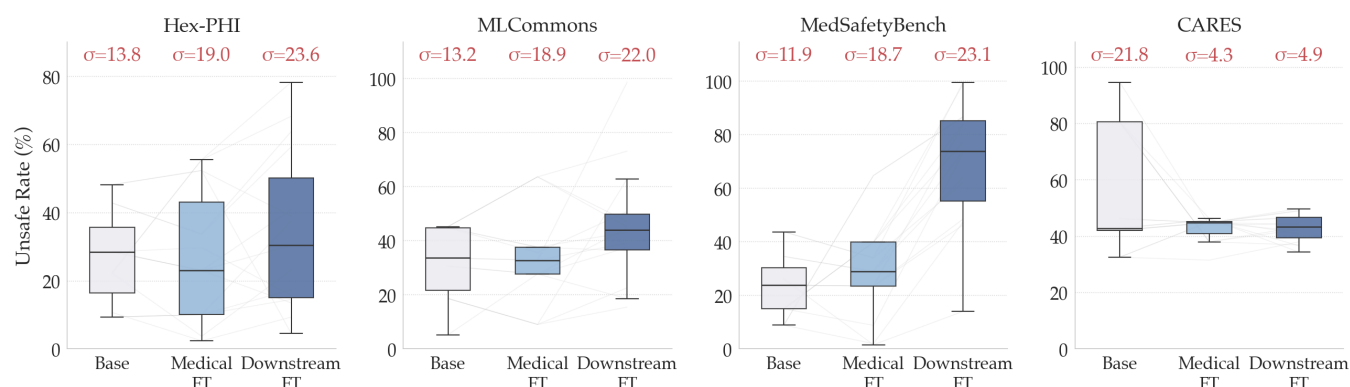


Figure 2: Safety distribution across fine-tuning lineages. Each box represents the interquartile range of safety drift at a given stage, with the median shown as a horizontal line and the σ value denoting the standard deviation of unsafe-response rates at that stage. Overall, safety behavior often becomes more variable across fine-tuning lineages.

at the second, and directional stability remained limited: between 60% and 86% of lineages exhibited sign reversals between stages, depending on the benchmark.

Overall, these results indicate that measures of model safety often become more variable and less predictable across fine-tuning lineages, though the magnitude and form of instability depend strongly on how safety is measured.

Legal Domain To test the generality of safety drift, we conducted a similar ecological analysis for legal applications. This setting highlights how infrastructure limitations—specifically, a less developed evaluation ecosystem—constrain the visibility of safety risks.

Like the medical domain, legal fine-tunes exhibited no consistent directional pattern compared to their base models. Considering only benchmarks that directly assess unsafe behavior (HEX-PHI, MLCommons, Trident), 60% of fine-tuned legal models showed mixed-sign drift, improving on at least one benchmark while worsening on another. When legal instruments capturing related constructs—refusal behavior (SORRY-Bench) and legal adherence (SafeLawBench)—were included, the mixed-sign rate rose

to 93%, indicating near-universal inconsistency when multiple safety dimensions are evaluated simultaneously.

While median drift remained modest ($\approx \pm 3$ percentage points), these aggregates obscured substantial model-level volatility (Figure 3). Most benchmarks, particularly SORRY-Bench and Trident, showed wide bidirectional swings, with models split nearly evenly between improvement and degradation (47% vs. 53%). In contrast, SafeLawBench skewed positively, with 73% of models moving to safer behavior. These patterns suggest that apparent safety trends depend strongly on the construct measured, motivating more granular cross-benchmark analysis.

Construct Mismatch To probe the source of this inconsistency, we examined agreement between legal benchmarks that measure distinct safety-related constructs. Unlike the medical domain, where benchmarks target largely overlapping notions of professional harm, legal evaluations target heterogeneous objectives, including avoidance of explicitly harmful outputs, refusal behavior in response to risky requests, and adherence to applicable legal rules.

Agreement was notably weak, with Spearman correlations between Trident, SORRY-Bench, and SafeLawBench

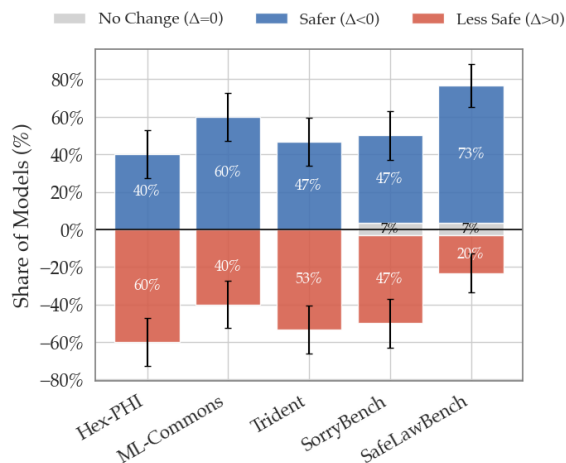


Figure 3: **Legal safety drift.** Share of analyzed models rated safer ($\Delta < 0$) vs. less safe ($\Delta > 0$) across legal benchmarks.

ranging from $\rho \approx 0.06$ – 0.17 . Fully coherent behavior, where all three benchmarks moved in the same direction, occurred in only 23% of models. These results suggest that legal fine-tuning can simultaneously improve one aspect of safety while degrading another, depending on the evaluative lens. Although disagreement among measures targeting different facets of safety is not inherently a problem, it raises questions about whether observed differences reflect limitations in reliability and validity or more substantive differences in construct operationalization. Because these measures lack external validation against ground-truth harm data, these explanations cannot be disentangled. Accordingly, inferences about legal model safety based on any single instrument should be regarded skeptically.

When viewed in tandem, both ecological analyses reveal that fine-tuning is frequently associated with benchmark-dependent and internally inconsistent changes in measured safety. Safety drift is often bidirectional, varies substantially across individual models, and becomes more variable across successive fine-tuning stages. Importantly, these patterns arise in real-world, non-adversarial development pipelines, underscoring that alignment instability is not confined to synthetic or malicious settings.

However, because the models we analyze differ in data, training procedures, and other often undocumented design choices, these results alone do not isolate the factors associated with observed changes in safety. Thus, to complement this ecological analysis, we conduct controlled fine-tuning experiments mirroring simple PEFT workflows. By fixing data and hyperparameters while varying fine-tuning method (LoRA, QLoRA, FFT) and base model family, these experiments allow us to assess whether patterns of safety drift persist when key aspects of the training setup are held constant, and whether common engineering choices provide reliable signals about post-fine-tuning safety behavior.

Phase II: Fine-tuning Experiments

Medical Domain Directional and Benchmark Inconsistency. Even under controlled conditions, fine-tuning produced highly divergent safety outcomes. On domain-specific medical benchmarks (MedSafetyBench, CARES), 83% of configurations became safer. Conversely, 100% degraded on the general-purpose MLCommons benchmark, while outcomes on HEx-PHI were nearly evenly split with 42% shifting unsafe. Median shifts ranged from a 12.2pp improvement (MedSafetyBench) to a 26.4pp degradation (MLCommons), indicating substantial redistribution of safety behavior rather than uniform improvement. One interpretation of this is that task-specific model optimization may promote more cautious behavior within explicitly trained medical contexts while simultaneously weakening it elsewhere, yielding uneven alignment across domains.

Consistent with the ecological analysis, every model exhibited mixed-sign drift, revealing trade-offs across benchmarks. This non-monotonic behavior persisted despite fixed data, hyperparameters, and training regimes, suggesting that fine-tuning design choices alone do not reliably determine the direction of safety change. Agreement across benchmarks also remained limited and sometimes contradictory, with correlations ranging from strongly positive (e.g., HEx-PHI vs. MedSafetyBench, $\rho \approx 0.9$) to strongly negative (e.g., CARES vs. MLCommons, $\rho \approx -0.6$). As in Phase I, these discrepancies indicate that measured “safety” depends strongly on evaluation design.

Training Regime Impact. Across benchmarks, we did not find strong evidence that any fine-tuning method consistently preserved alignment. Safety outcomes varied substantially by benchmark and configuration, with the same method producing opposing effects. For example, QLoRA improved Gemma substantially on CARES (-36 pp) while sharply degrading it on MLCommons ($+45$ pp). Full fine-tuning produced the single largest improvement observed (Gemma, -69 pp on CARES) but showed negligible or negative changes on other benchmarks. Similar inconsistencies appeared across LoRA and QLoRA configurations. Base model family was likewise an unreliable predictor: although Gemma achieved large improvements on CARES (-69.2 pp), it also exhibited pronounced degradation on MLCommons ($+33$ – 45 pp). Across all models and methods, MLCommons consistently showed higher residual harmfulness after fine-tuning—a concerning result given that it is the most systematically developed benchmark in our suite, with a dedicated classifier aligned to an explicit hazard taxonomy rather than an off-the-shelf LLM judge.

Proxy Failure. A natural hypothesis is that larger parameter updates induce greater behavior drift. If true, the magnitude of weight-space movement during fine-tuning should correlate with safety changes. To test this, we measured the normalized L_2 distance between each fine-tune and its base as a proxy for parameter change. Across benchmarks, no systematic relationship emerged: small updates sometimes produced large safety shifts and larger updates often yielded minimal or opposing effects ($|\rho| < 0.25$, $R^2 < 0.1$).

On MedSafetyBench, (Figure 5(a)), all three Mistral variants (LoRA, QLoRA, FFT) achieved similar safety gains

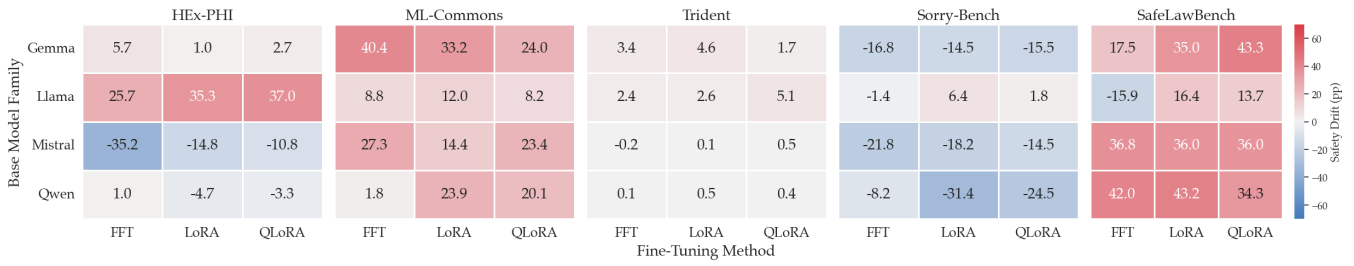


Figure 4: Safety drift (Δ unsafe rate, pp) post legal fine-tuning segmented by model, fine-tuning method, and benchmark. Safety effects vary by base model and method, with no tuning approach consistently preserving alignment.

(≈ -35 to -37 pp) despite parameter distances spanning nearly an order of magnitude (≈ 0.01 – 0.29). Conversely, Gemma–QLoRA and Mistral–LoRA exhibited opposite safety outcomes despite comparable parameter movement (≈ 0.01). These patterns indicate that parameter distance is a poor proxy for alignment change: minimal updates can induce substantial behavioral shifts, while large updates may leave safety largely unchanged.

We next examine whether similar limitations hold in the legal domain, where safety evaluations target different constructs and exhibit greater measurement heterogeneity.

Legal Domain Construct-Specific Drift. Controlled legal fine-tuning similarly produced construct-dependent safety shifts rather than uniform improvement. Across all base models and fine-tuning methods, MLCommons consistently degraded, with especially large increases in unsafe behavior for Gemma and Mistral under FFT and QLoRA (often exceeding $+30$ pp, mean $\Delta = +19.8$ pp). SafeLawBench also exhibited widespread degradation (92%), particularly for Mistral- and Qwen-based models. Conversely, SORRY-Bench frequently improved, while HEX-PHI displayed method-dependent sign reversals.

Pervasive mixed-sign drift was observed across models, with 83% improving under one evaluation while worsening under another. Restricting analysis to benchmarks directly assessing unsafe behavior (HEX-PHI, MLCommons, Trident) reduced the mixed-sign rate to 42%, indicating that construct mismatch substantially amplified—but did not fully account for—directional instability.

The persistence of this trend under controlled conditions across both domain settings suggests that non-monotonic alignment changes are a general feature of domain adaptation rather than an artifact of noisy training.

Evaluation Disagreement and Construct Mismatch

Agreement across legal evaluation instruments was limited. Rank correlations between benchmark-specific safety changes varied widely, from strongly positive (e.g., HEX-PHI vs. Trident, $\rho = +0.82$) to strongly negative (e.g., HEX-PHI vs. SafeLawBench, $\rho = -0.63$; MLCommons vs. SORRY-Bench, $\rho = -0.55$). Some disagreement involving SafeLawBench seemed to stem from evaluation–objective mismatch: its template favored concise, reference-style answers, whereas legal fine-tuning often produced more contextualized explanations. Behavior plausibly reflecting improved legal reasoning may have been penalized, highlight-

ing the sensitivity of measured safety to evaluation criteria.

Proxy Failure. Mirroring the medical results, no reliable relationship emerged between normalized parameter distance (L_2) and changes in unsafe behavior on MLCommons or Trident. However, HEX-PHI exhibited a significant negative correlation ($\rho < -0.64$, $p = 0.02$), suggesting that smaller, more targeted updates, typical of PEFT, can sometimes be associated with greater safety degradation as opposed to larger structural changes (Figure 5(b)).

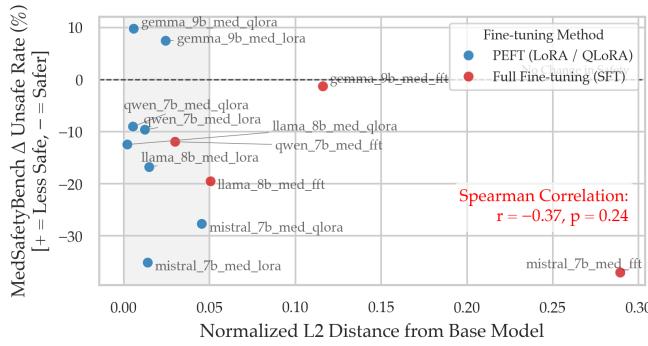
Across both domains, controlled experiments confirm that safety drift persists even when data, hyperparameters, and training regimes are held constant. Fine-tuning method, base-model family, and tuning magnitude individually offer limited predictive power over post-fine-tuning safety outcomes. That these patterns emerge under controlled conditions reinforces the ecological finding that safety drift is a general property of domain adaptation rather than an artifact of heterogeneous training pipelines.

Qualitative Patterns

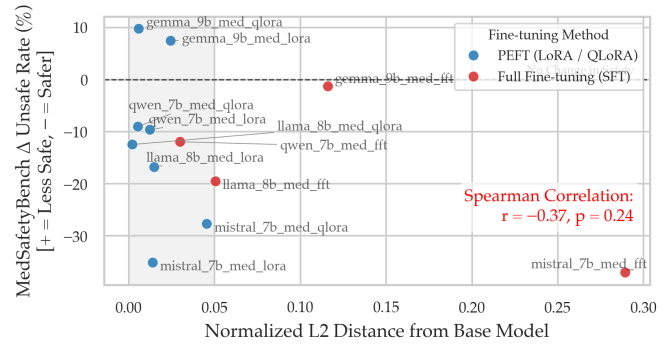
Examination of high-scoring harmful prompt–response pairs revealed recurring failure modes across domains. Most prominently, fine-tuned models showed increased willingness to comply with harmful requests when framed in domain-appropriate language—generating detailed self-harm guidance, unsolicited diagnoses with specific medication recommendations, or strategic advice on introducing unsubstantiated legal allegations. Fine-tuned models were also more likely to produce complete, realistic written artifacts (clinical plans, legal strategies, social media posts) that could cause direct harm if acted upon. Medical fine-tunes in particular exhibited *persona fixation*, defaulting to a clinical role even when contextually inappropriate, interpreting unrelated requests through a medical lens or producing garbled outputs. Whether degraded out-of-domain responses reflect desirable scope narrowing or unpredictable reliability loss requires further analysis, but both carry implications for auditability and trust in high-stakes deployment.

Evaluation Modification Experiment

Because safety outcomes varied sharply across benchmarks, we examined whether modifying the evaluation setup systematically affected the observed relationship between base and fine-tuned models. Specifically, we augmented the HEX-PHI judging template with example responses illustrating

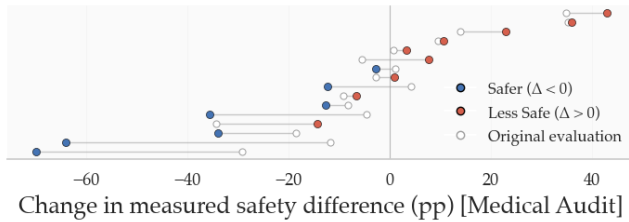


(a) Medical fine-tunes (MedSafetyBench).

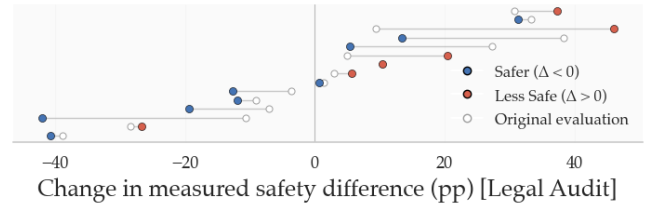


(b) Legal fine-tunes (HEX-PHI).

Figure 5: Safety change vs. parameter shift under tuning. On MedSafetyBench, safety effects diverge sharply among minimally modified models ($L_2 < 0.05$), while in law, smaller updates can be associated with large safety change ($\rho < -0.64$, $p = 0.02$).



(a): Medical Model Analysis



(b): Legal Model Analysis

Figure 6: Sensitivity to evaluation setup. Across both analyses, after modifying the judging template, safety shifts (pp) are bidirectional, indicating that evaluation framing can influence estimated alignment effects as strongly as fine-tuning itself.

each point on the 1–5 safety scale. In principle, providing concrete examples should give the judge clearer guidance on how to apply the rubric, reducing judgment noise and yielding more stable comparisons. If fine-tuned models were genuinely safer—or less safe—we would expect these differences to emerge consistently, and potentially more strongly, under clearer evaluative standards.

Contrary to this expectation, no systematic stabilization emerged. As shown in Figure 6, adding examples sometimes widened the apparent safety gap in favor of the fine-tuned model, amplified it in the opposite direction or produced minimal change. Models also exhibited bidirectional shifts, with 25% of medical results reversing sign under the modified template. These patterns suggest that evaluation-setup choices influence estimated alignment effects as much as fine-tuning itself, raising fundamental questions about how safety differences should be interpreted when the evaluation instrument introduces substantial measurement variance.

Policy and Governance Implications

Our study shows that benign, domain-specific fine-tuning can induce large and difficult-to-predict changes in model safety assessments, even when the adaptation objective is unrelated to the safety construct being evaluated. Across both phases and domains, differences in the safety of base and fine-tuned models were frequently bidirectional and heterogeneous. Together, our findings support four core claims.

Model safety can be affected by benign, task-specific fine-tuning, not just by adversarial approaches. While prior work has emphasized adversarial fine-tuning as an attack vector, our results point to a more pervasive source of risk: ordinary, capability-driven adaptation that unintentionally perturbs safety behavior. The cumulative effects of routine fine-tuning across the supply chain may pose a larger risk than rare malicious interventions. This frames the central challenge as one of supply-chain reliability: whether downstream developers can apply common adaptation methods without degrading safety, and whether upstream providers can offer meaningful guarantees or support for managing risks that emerge under routine fine-tuning.

Safety drift is an emergent property of the model lifecycle, requiring shared responsibility across the supply chain. Our findings highlight an accountability gap. Upstream model providers control general-purpose safety mechanisms but cannot anticipate the full range of risks that arise once models are adapted for specialized tasks. Downstream developers, by contrast, are best positioned to identify task-specific risks, yet often lack the resources or expertise to manage harms beyond their intended use cases. At first glance, it may seem reasonable to assign responsibility to downstream deployers since fine-tuning modifies the upstream model and alters its safety profile. However, our results complicate this view. Given that even routine fine-tuning choices can unpredictably degrade safety in ways that

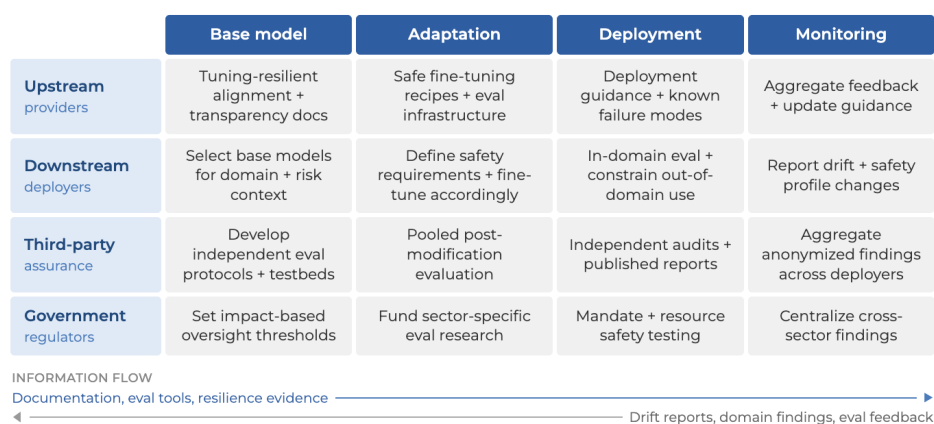


Figure 7: Lifecycle-aware governance framework mapping supply-chain actors to responsibilities with bidirectional information flow: upstream documentation enabling informed adaptation, downstream reporting enabling collective improvement

may not be reliably inferred, there seems to be no easy way to integrate a general-purpose model into a specialized system without risking unintended safety degradation.

A useful analogy clarifies why deployer-only responsibility is insufficient. Suppose an engine manufacturer supplies configurable engines to vehicle makers, who integrate them into vehicles built for various operating conditions. The engines come with configurable components (e.g., turbocharger settings, fuel injection systems, electronic control units) that vehicle manufacturers can tune for their use cases. If routine adjustments caused unpredictable safety failures, responsibility would not rest solely with the vehicle manufacturer; the engine supplier would also bear responsibility for ensuring reliable behavior under foreseeable integration. By analogy, if foundation models function as shared “engines” and fine-tuners as “manufacturers” assembling domain-specific systems, meaningful safety governance must span both upstream and downstream actors.

This principle is reinforced by substantial resource and information asymmetries. Frontier AI companies are projected to spend over \$500 billion on capital expenditures (Goldman Sachs 2025), while downstream deployers often operate as solo developers fine-tuning on consumer-grade hardware. Model providers also possess critical information unavailable to downstream developers, like which harmful behaviors were targeted, what alignment data was used, and which safety techniques were applied. Without these details, developers cannot know which safety properties they may inadvertently degrade during fine-tuning. These asymmetries are further amplified in closed API systems that withhold training details, model weights, and internal evaluations.

The result is a structural assurance gap: responsibility for comprehensive safety evaluation risks falling on the least-informed actors: those closest to deployment but furthest from the resources needed to validate safety. Effective governance should therefore treat safety drift as an emergent property of the full deployment pipeline, arising from interactions among the base model, fine-tuning data, and adaptation methods, and account for AI supply-chain asymmetries.

Responsibility allocation should thus reflect both positioning and capacity (Figure 7). Upstream providers, with their complete access and substantial resources, are best positioned to develop robust alignment, provide evaluation infrastructure, and publish tuning-resilience evidence and known failure modes. Downstream deployers bring domain expertise and deployment proximity, making them best suited for context-specific evaluation, scope management, and reporting modification effects. Third-party assurance organizations can bridge gaps by pooling evaluation resources and aggregating anonymized findings across deployers. Government agencies can set impact-based oversight standards and facilitate cross-sector coordination where market incentives prevent sharing. All in all, effective lifecycle governance seems to require structured information flows in both directions across the supply chain.

Policy thresholds tied to modification magnitude are unlikely to be effective. Our results show that even small, benign parameter changes can produce substantial safety shifts, while more extensive adaptation can sometimes improve safety. What matters is not the magnitude of the technical change but its behavioral impact—a distinction that current governance frameworks largely fail to capture.

Under current European Commission guidance, a downstream modifier becomes a new provider of a general-purpose AI model if their adaptation compute exceeds one-third of the original training compute, triggering stricter legal obligations including safety re-assessment duties (European Commission 2025). California’s SB 53 leaves frontier developers to define “substantial modification” but also anchors its definition of frontier models to compute thresholds, encouraging compute-based interpretations (California State Legislature 2025). These approaches risk governing based on resource expense instead of behavioral impact, missing safety-relevant tuning that falls below their thresholds.

Colorado’s Consumer Protections for Artificial Intelligence Act (SB 24-205) offers a contrasting approach, defining “substantial modification” as a deliberate change that results in new reasonably foreseeable risks—focusing on be-

havioral impact rather than modification magnitude (Colorado General Assembly 2024). This framing better reflects the uncertainty we document. Still, operationalizing impact-based definitions requires robust evaluation approaches for modified models, which remain underdeveloped.

Measurement instability complicates the interpretation of safety drift. For safety drift to inform governance, the measurements used to detect it must be both valid and reliable. We identify three critical challenges.

1. *Divergent benchmark results obscure clear guidance for deployment and limit reliable risk assessment.* Evaluations across multiple benchmarks that nominally measure similar constructs sometimes yielded conflicting results, with weak rank correlations. This may reflect legitimate differences in construct operationalization or may indicate underlying reliability or validity limitations. Without empirical validation against expert judgment or real-world harm data, these explanations cannot be disentangled. Preliminary “smell tests” (Blodgett et al. 2021) can identify obvious conceptual flaws, but the field has yet to develop rigorous methods for interpreting benchmark disagreement (Bean et al. 2025). Ultimately, neither the direction nor the magnitude of safety drift observed can be assumed to generalize across alternative definitions or instruments; many benchmarks may capture compliance with their specific operationalization rather than ecologically meaningful safety.
2. *General-purpose safety definitions may conflict with domain-specific goals.* Refusal to provide legal guidance may be appropriate for a general-purpose model but constitutes non-performance for a legal assistant. We observed this mismatch with SORRY-Bench (Xie et al. 2025), which rewarded refusal to answer legal queries, thus penalizing the utility that legal fine-tuning aimed to provide. In these cases, generalist metrics may lack validity in specialized settings, measuring resistance to task execution rather than the safety of task performance. This calls for domain-specific evaluations explicitly designed for post-modification contexts, validated to ensure their construct definitions remain appropriate after adaptation.
3. *Seemingly minor implementation choices in LLM-based evaluations materially impact safety estimates.* Adding explanatory examples to evaluation prompts while holding outputs fixed substantially changed measured outcomes in our experiments. This suggests that some apparent safety drift may stem from interactions between fine-tuning-induced stylistic changes and evaluation protocols rather than substantive behavioral change. When safety estimates depend on judge design or setup, evaluation reliability constrains accurate interpretation.

These challenges point to a broader infrastructure need. The first step in managing safety drift is measuring it, which requires both general-purpose and domain-specific evaluations explicitly designed with post-modification use in mind. Valid evaluations must be supported by accessible infrastructure like code libraries, testbeds, and platforms that reduce barriers for resource-constrained developers. Current infrastructure concentrates at two extremes: general-purpose

frameworks requiring significant compute and technical expertise, and domain-specific platforms developed by well-resourced specialized communities. This leaves a gap for developers in less common domains who need validated evaluation tools accounting for post-modification behavior but lack resources to develop them independently.

Closing this gap requires shared benchmarking infrastructure maintained independently of evaluated entities, standardized reporting formats, and bidirectional documentation flows between actors. This mirrors established practices in cybersecurity, where coordinated vulnerability disclosure enables parallel mitigation development (Cattell, Ghosh, and Kaffee 2024; Householder et al. 2025).

Future Research Directions. Based on these findings, there is value in pursuing the development of deeper, tuning-resilient alignment and domain-scoping methods that remain practical for resource-constrained deployer (Qi et al. 2024; Alssum et al. 2025; Yang et al. 2025; Shi et al. 2025). Future work should also test whether similar drift occurs in closed-API fine-tuning, where opaque training details and evaluations may intensify information asymmetries (Thinking Machines Lab 2026).

Limitations

This study characterizes safety changes across real-world fine-tuned model ecosystems rather than causal mechanisms. Our ecological comparisons therefore capture real-world associations between fine-tuning and safety drift, not the effects of specific design choices. Phase II partially addresses this by holding data, hyperparameters, and adaptation methods fixed, though across a narrower set of models. Our results also depend on the benchmarks and judges used. Systemic bias is mitigated through the use of paired analysis capturing relative drift rather than absolute safety level though broader sensitivity analysis across alternative evaluation setups remains future work. Finally, we study English-language open-weight models in the 7–70B range so generalization to closed APIs, multilingual/frontier models, and other domains remains open. Still, the recurrence of drift across model families, domains, benchmarks, and tuning methods suggests these dynamics are not architecture- or configuration-specific.

Conclusion

Our work challenges the assumption that baseline safety evaluations remain stable under non-adversarial downstream fine-tuning. Across ecological and controlled analyses in medicine and law, we show that benign adaptation produces large, heterogeneous, and benchmark-dependent safety shifts that cannot be reliably inherited from base models or predicted from modification scale. We therefore argue for supply-chain aware governance in which upstream providers publish tuning-resilience evidence, downstream deployers conduct domain-grounded post-modification evaluation, third parties maintain shared benchmarking infrastructure, and regulators emphasize consequence-based rather than magnitude-based oversight.

References

- Alssum, L.; Itani, H.; Hammoud, H. A. A. K.; Torr, P.; Bibi, A.; and Ghanem, B. 2025. Unforgotten Safety: Preserving Safety Alignment of Large Language Models with Continual Learning. *arXiv:2512.10150*.
- Bean, A. M.; Kearns, R. O.; Romanou, A.; Hafner, F. S.; Mayne, H.; Batzner, J.; Foroutan, N.; Schmitz, C.; Korgul, K.; Batra, H.; Deb, O.; Beharry, E.; Emde, C.; Foster, T.; Gausen, A.; Grandury, M.; Han, S.; Hofmann, V.; Ibrahim, L.; Kim, H.; Kirk, H. R.; Lin, F.; Liu, G. K.-M.; Luettgau, L.; Magomere, J.; Rysström, J.; Sotnikova, A.; Yang, Y.; Zhao, Y.; Bibi, A.; Bosselut, A.; Clark, R.; Cohan, A.; Foerster, J.; Gal, Y.; Hale, S. A.; Raji, I. D.; Summerfield, C.; Torr, P. H. S.; Ududec, C.; Rocher, L.; and Mahdi, A. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. *arXiv:2511.04703*.
- Betley, J.; Tan, D.; Warncke, N.; Szyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv:2502.17424*.
- Bhardwaj, R.; and Poria, S. 2023. Language model unalignment: Parametric red-teaming to expose hidden harms and biases. *arXiv preprint arXiv:2310.14303*.
- Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015. Online: Association for Computational Linguistics.
- Bommasani, R.; Singer, S. R.; Appel, R. E.; Cen, S.; Cooper, A. F.; Cryst, E.; Gailmard, L. A.; Klaus, I.; Lee, M. M.; Raji, I. D.; Reuel, A.; Spence, D.; Wan, A.; Wang, A.; Zhang, D.; Ho, D. E.; Liang, P.; Song, D.; Gonzalez, J. E.; Zittrain, J.; Chayes, J. T.; Cuéllar, M.-F.; and Fei-Fei, L. 2025. The California Report on Frontier AI Policy. Published June 17, 2025.
- California State Legislature. 2025. Senate Bill No. 53: Artificial intelligence models: large developers. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=20250260SB53. 2025–2026 Regular Session, accessed 2026-05-20.
- Cao, C.; Zhu, H.; Ji, J.; Sun, Q.; Zhu, Z.; Yinyu, W.; Dai, J.; Yang, Y.; Han, S.; and Guo, Y. 2025. SafeLawBench: Towards Safe Alignment of Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 14015–14048. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Cattell, S.; Ghosh, A.; and Kaffee, L.-A. 2024. Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities. *arXiv preprint arXiv:2402.07039*.
- Chen, S.; Li, X.; Zhang, M.; Jiang, E. H.; Zeng, Q.; and Yu, C.-H. 2025. CARES: Comprehensive Evaluation of Safety and Adversarial Robustness in Medical LLMs. *arXiv:2505.11413*.
- Colorado General Assembly. 2024. SB24-205: Consumer Protections for Artificial Intelligence. <https://leg.colorado.gov/bills/sb24-205>. Regular Session, signed into law May 17, 2024; accessed 2026-05-20.
- Couture, R. J. 2025. The Impact of Artificial Intelligence on Law Firms’ Business Models. *Harvard Law School Center on the Legal Profession, Insights*. Qualitative study of AI adoption and business models in AmLaw 100 firms.
- Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv:2305.14314*.
- Draper, C.; and Gillibrand, N. 2023. The Potential for Jurisdictional Challenges to AI or LLM Training Datasets. In *AI4AJ@ICAIL*.
- Endal, O.; Vrcic, A.; Nag, S.; Malter, N.; and Araz, D. 2025. Modifying AI Under the EU AI Act: Lessons from Practice on Classification and Compliance.
- European Commission. 2025. Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act. <https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>. Accessed: September 24, 2026.
- European Parliament and Council of the European Union. 2024. Article 11: Technical Documentation. In *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, EU 2024/1689. Official Journal of the European Union. Accessed: 2025-10-27.
- Fraser, K. C.; Dawkins, H.; Nejadgholi, I.; and Kiritchenko, S. 2025. Fine-Tuning Lowers Safety and Disrupts Evaluation Consistency. In Derczynski, L.; Novikova, J.; and Chen, M., eds., *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, 129–141. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-279-4.
- Goldman Sachs. 2025. Why AI Companies May Invest More than \$500 Billion in 2026. Accessed: 2026-05-20.
- Han, T.; Kumar, A.; Agarwal, C.; and Lakkaraju, H. 2024. MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 33423–33454. Curran Associates, Inc.
- Hawkins, W.; Mittelstadt, B.; and Russell, C. 2024. The effect of fine-tuning on language model toxicity. *arXiv preprint arXiv:2410.15821*.
- He, L.; Xia, M.; and Henderson, P. 2024a. What is in Your Safe Data? Identifying Benign Data that Breaks Safety. *arXiv:2404.01099*.
- He, L.; Xia, M.; and Henderson, P. 2024b. What’s in Your “Safe” Data?: Identifying Benign Data that Breaks Safety. *arXiv preprint, arXiv:2404.01099*.
- Henderson, P.; et al. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv preprint arXiv:2308.11462*.

- Householder, A. D.; Sarvepalli, V. S.; Havrilla, J.; Churilla, M.; Pons, L.; hon Lau, S.; VanHoudnos, N. M.; Kompanek, A.; and McIlvenny, L. 2025. Protecting AI from the Outside In: The Case for Coordinated Vulnerability Disclosure.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*.
- Hugging Face. 2023a. Parameter-Efficient Fine-Tuning (PEFT) for Large Language Models. <https://huggingface.co/blog/peft>. Blog post introducing the PEFT library and its support for methods like LoRA and QLoRA.
- Hugging Face. 2023b. PEFT: Parameter-Efficient Fine-Tuning of Transformers. <https://github.com/huggingface/peft>. Python library supporting LoRA, QLoRA, and related PEFT methods.
- Jain, S.; Lubana, E. S.; Oksuz, K.; Joy, T.; Torr, P. H. S.; Sanyal, A.; and Dokania, P. K. 2024. What Makes and Breaks Safety Fine-tuning? A Mechanistic Study. *arXiv:2407.10264*.
- Kumar, D.; Kumar, A.; Agarwal, S.; and Harshangi, P. 2024. Increased LLM Vulnerabilities from Fine-Tuning and Quantization. *arXiv preprint*, *arXiv:2404.04392*.
- Ladhak, F.; Durmus, E.; Suzgun, M.; Zhang, T.; Jurafsky, D.; McKeown, K.; and Hashimoto, T. B. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3206–3219.
- Laufer, B.; Oderinwale, H.; and Kleinberg, J. 2025. Anatomy of a Machine Learning Ecosystem: 2 Million Models on Hugging Face. *arXiv preprint* *arXiv:2508.06811*.
- Lehmann, J. S.; et al. 2025. Implementing large language models in healthcare while balancing innovation, privacy, and safety. *PMC / NPJ Digital Medicine*, PMC11885444.
- Lermen, S.; Rogers-Smith, C.; and Ladish, J. 2023. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. *arXiv preprint*, *arXiv:2310.20624*.
- LexisNexis. 2024. Survey Reveals How Gen AI is Reshaping Law. Global legal AI adoption and investment survey.
- Liu, L.; Yang, X.; Lei, J.; Shen, Y.; Wang, J.; Wei, P.; Chu, Z.; Qin, Z.; and Ren, K. 2024. A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions. *arXiv:2406.03712*.
- LLama, M. 2023. meta-llama/llama-cookbook.
- MLCommons. 2025. AILuminate - MLCommons — [mlcommons.org](https://mlcommons.org/benchmarks/ailuminate/). <https://mlcommons.org/benchmarks/ailuminate/>. [Accessed 27-10-2025].
- Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *arXiv:2406.05946*.
- Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv:2310.03693*.
- Retool. 2024. Retool State of AI 2024 Report: How People Actually Use AI. <https://codingscape.com/blog/retool-state-of-ai-2024-report-how-people-actually-use-ai>.
- Richter, A. 2025. Self-Hosting AI: For Privacy, Compliance, and Cost Efficiency. <https://techgdpr.com/blog/self-hosting-ai-for-privacy-compliance-and-cost-efficiency/>. Accessed: 2026-01-04.
- Salman, H.; Jain, S.; Ilyas, A.; Engstrom, L.; Wong, E.; and Madry, A. 2022. When does bias transfer in transfer learning? *arXiv preprint* *arXiv:2207.02842*.
- See, Y. K. C.; Lim, K. S. A.; Au, W. Y.; Chia, S. Y. C.; Fan, X.; and Li, Z. K. 2025. The use of Large Language Models in ophthalmology: A scoping review on current use-cases and considerations for future works in this field. *Big Data Cogn. Comput.*, 9(6): 151.
- Shi, G.; Lu, Z.; Dong, X.; Zhang, W.; Zhang, X.; Feng, Y.; and Wu, X.-M. 2025. Understanding Layer Significance in LLM Alignment. *arXiv:2410.17875*.
- Sun, A. Y.; Zemor, E.; Saxena, A.; Vaidyanathan, U.; Lin, E.; Lau, C.; and Mugunthan, V. 2023. Does Fine-Tuning GPT-3 with the OpenAI API Leak Personally-Identifiable Information? *arXiv preprint*, *arXiv:2307.16382*.
- Thinking Machines Lab. 2026. Tinker: A Training API for Researchers. <https://thinkingmachines.ai/tinker/>.
- Thomson Reuters. 2025. 2025 Generative AI in Professional Services Report. Technical report, Thomson Reuters. Accessed: 2026-01-04.
- U.S. Department of Health and Human Services, Office for Civil Rights. 2025. HIPAA Security Rule To Strengthen the Cybersecurity of Electronic Protected Health Information. Federal Register, Proposed Rule (90 FR 898). Doc. No. 2024-30983; RIN 0945-AA22; Comments due March 7, 2025.
- Wan, A.; Wallace, E.; Shen, S.; and Klein, D. 2023. Poisoning Language Models During Instruction Tuning. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 35413–35425. PMLR.
- Wang, A.; and Russakovsky, O. 2023. Overwriting pre-trained bias with finetuning data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3957–3968.
- Wang, X.; Zhang, N. X.; He, H.; Nguyen, T.; Yu, K.-H.; Deng, H.; Brandt, C.; Bitterman, D. S.; Pan, L.; Cheng, C.-Y.; Zou, J.; and Liu, D. 2024. Safety challenges of AI in medicine in the era of large language models.
- Wei, B.; Huang, K.; Huang, Y.; Xie, T.; Qi, X.; Xia, M.; Mittal, P.; Wang, M.; and Henderson, P. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint* *arXiv:2402.05162*.
- Winecoff, A.; and Bogen, M. 2025. Improving governance outcomes through AI documentation: Bridging theory and practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Xie, T.; Qi, X.; Zeng, Y.; Huang, Y.; Sehwag, U. M.; Huang, K.; He, L.; Wei, B.; Li, D.; Sheng, Y.; Jia, R.; Li, B.; Li, K.; Chen, D.; Henderson, P.; and Mittal, P. 2025. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Yang, K.; Tao, G.; Chen, X.; and Xu, J. 2025. Alleviating the Fear of Losing Alignment in LLM Fine-tuning. arXiv:2504.09757.

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

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