A GRAVITATIONAL INTERPRETATION OF SAFETY REVERSION UNDER FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment in large language models can degrade during post-training even when neither the data nor the objective is intentionally adversarial. Alignment rebound and reverse dynamics suggest that this degradation may reactivate behavior suppressed during safety alignment. Building on these ideas, we hypothesize that ordinary non-adversarial post-training follows a *reversion direction*: the activation-space displacement from the safety-aligned model toward a more permissive, earlier helpful-only state. We see that for Llama, every tested trajectory across references, tasks, and seeds exceeds a matched empirical null, while at aligned Llama and Qwen checkpoints, a vocabulary readout shows that the direction locally favors task-engaging over fixed refusal-like openings. Its geometric expression is behaviorally informative: as post-training proceeds, alignment with the direction and harmfulness increase together, yielding a strong descriptive correlation (Spearman $r = 0.958$). To move beyond correlation, we test causal relevance during adaptation using objectives constructed from this coordinate. Across all tested Llama, Qwen, and Gemma settings from 3B to 14B, an optimizer-matched objective opposing positive motion reduces geometric alignment and harmfulness relative to ordinary fine-tuning, whereas a separately stabilized objective reinforcing that motion increases both. Every model and scale exhibits the same mean block–baseline–push ordering, showing that the causal relevance of the reversion direction is not tied to one architecture or model size. Finally, we show that a standard safety-rehearsal objective, built without access to the direction, independently opposes it and cuts cumulative reversion by about 30% in Llama and Qwen.

1 INTRODUCTION

After broad pretraining, language models are commonly post-trained not only to follow instructions (Wei et al., 2022; Ouyang et al., 2022) but also to refuse or safely redirect requests for harmful assistance (*i.e. safety alignment*) (Bai et al., 2022). However, a safety-aligned model is rarely the endpoint of training: it is often specialized further through gradient-based updates for a new task or domain (Gururangan et al., 2020; Hu et al., 2022; Horwitz et al., 2025). This later adaptation is called *benign* when its downstream examples and targets are intentionally non-harmful and its task loss does not reward harmful assistance (He et al., 2024). Crucially, the term *benign* characterizes the supervision rather than guaranteeing the behavior of the resulting model, because even fine-tuning on such *benign* data can still substantially increase compliance with harmful requests (Qi et al., 2024); even among benign examples, subsets selected by safety-relevant criteria can cause much greater degradation than random subsets (He et al., 2024; Guan et al., 2025; Alam et al., 2026; Poppi et al., 2025). Safety acquired during post-training is therefore not automatically preserved under later optimization.

Understanding this failure requires more than establishing that safety is fragile: we must also characterize the dynamics through which harmful behavior can re-emerge under benign adaptation. Existing work shows that refusal can depend disproportionately on early output tokens, that sparse or low-rank changes can compromise safety, and that measured safety remains stable only within a limited neighborhood of an aligned checkpoint (Qi et al., 2025; Wei et al., 2024; Peng et al., 2024), and that narrow fine-tuning can induce broad misalignment associated with a convergent linear representation (Betley et al., 2025; Soligo et al., 2025). While these findings identify mechanisms of

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

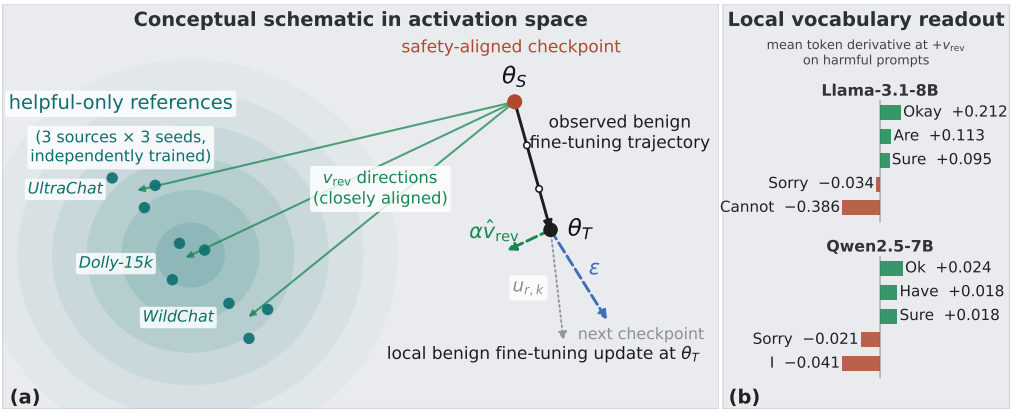


Figure 1: **Conceptual schematic of the gravitational interpretation**. **(a)** The hypothesized earlier, more permissive state is proxied by a family of independently trained helpful-only references (three sources, three seeds), *fixed before the prospective target trajectories are trained*. Benign fine-tuning yields task-specific changes (ϵ) plus a hypothesized reversion component ($\alpha\hat{v}_{\text{rev}}$); arrows are conceptual, not a training lineage. **(b)** Real mean token derivatives at $+v_{\text{rev}}$ for `Llama-3.1-8B-Instruct` and `Qwen2.5-7B-Instruct` on held-out harmful prompts (a local readout, not generated text), relative to each model's group mean: all shown tokens rise, red ones less than average.

susceptibility, complementary accounts of alignment elasticity and return dynamics across unsafe, task-agnostic, and selected benign adaptation regimes suggest that training history may also give subsequent degradation a systematic direction (Ji et al., 2025; Huang et al., 2026; Bach et al., 2026); nor is the phenomenon restricted to language models, since benign fine-tuning likewise erodes safety alignment and revives suppressed concepts in text-to-image diffusion models (Kim et al., 2024; Alam et al., 2026). Those accounts measure the return through output distributions relative to the base model or through competing terms in the training loss. We instead cast it as a single activation-space direction, pointing from the safety-aligned model toward independently built helpful-only states rather than toward the base model itself, and fixed before the prospective target trajectories are trained, so that ordinary benign adaptation can be tested against it and training objectives can block or amplify it directly. This motivates a reversion hypothesis for the benign regimes studied here.

Concretely, we hypothesize that part of this degradation is a *reversion* toward an earlier, more permissive state the model passed through before it was ever safety-aligned. A base pretrained checkpoint is not yet a chatbot, so to approximate that earlier state we lightly fine-tune the same base model into a *helpful-only* proxy, and define the reversion direction as the displacement from the safety-aligned model toward this proxy in activation space, where linear directions are a common description of model behavior (Park et al., 2024; Arditi et al., 2024). Because the proxy is built this way, defining it never touches how the target models actually behave later, which lets us ask, cleanly, whether ordinary benign adaptation drifts toward it more than a matched geometric baseline would predict. It does: once we fixed the reference family and the test in advance, every additional trajectory we tried moved in the predicted direction more than its matched null, with motion along the shared direction accounting for 29.5% of the observed activation-space path against a matched-null 95th percentile of just 6.8%.

As ordinary benign adaptation proceeds, we see alignment with the reversion direction and harmful behavior rise together, a strong descriptive correlation (Spearman $r = 0.958$). This raises a natural question: what, locally, does moving along this direction actually mean?

To understand the nature of this direction, we ask what moving along it locally promotes. Rather than a narrow rise in agreeable phrasing, its dominant effect is relaxing the refusal-template lock on the very first output token: on harmful prompts disjoint from direction construction, the canonical direction reopens a broad, ordinary distribution of response openers, and for both it and the inde-

pendently constructed family centroid, affirmative, task-engaging openings are significantly over-represented relative to a fixed refusal-like set beyond matched prompt-variability nulls. Measuring the same helpful-only–versus–aligned checkpoint differences on benign probes recovers related but non-identical directions, indicating a partly prompt-general request-engagement component together with harmful-probe-dependent structure. Both the correlation and this local signature are only associative, though, so to test whether the direction actually matters causally, we next intervene during training with objectives built directly from it.

Concretely, we build two such objectives: one that pushes against the direction (block) and one that reinforces it (push), each matched to its model's own optimizer settings across Llama, Qwen, and Gemma from 3B to 14B. The ordering is consistent across all of our experiments: mean harmfulness is lowest under block, higher under ordinary fine-tuning, and highest under push. As one illustrative point in this range, block brings harmfulness at step 100 down from $24.6\% \pm 8.3\%$ to $5.7\% \pm 1.9\%$ when fine-tuning `Llama-3.1-8B-Instruct`, close to the 2.8% rate of the safety-aligned anchor itself, while push raises it to $48.8\% \pm 4.3\%$. Crucially, this improvement does not come at a hidden cost: relative to ordinary fine-tuning, block adds no aggregate degradation on held-out Alpaca negative log-likelihood or IFEval in this setting, keeps IFEval comparable or higher in five of the six settings, and lowers mean harmfulness at every late layer we tested.

That said, we do not think v_{rev} is the only direction that can identify a safety-compliant checkpoint: some calibrated random-axis blocks help too, just inconsistently and with no principled way to pick a favorable axis ahead of time. What sets v_{rev} apart is that ① it is fixed, ② its intervention sign is predicted before we ever look at outcomes, and ③ it happens to be the direction that ordinary benign adaptation follows on its own.

However, many techniques already exist to preserve safety alignment during post-training adaptation (Hsu et al., 2024; Bhardwaj et al., 2024; Huang et al., 2024; 2025; Tamirisa et al., 2025), so, as a further independent check, we also test a safety-rehearsal loss term, one common way of preventing safety degradation in practice, that is not derived from v_{rev} at all. Its updates nonetheless naturally oppose the reversion direction at every tested Llama and Qwen state, and under the joint task-plus-retaining recipe, cumulative reversion falls by 30.9% and 30.3% while mean unsafe rates drop on both BeaverTails and HarmBench, evidence that the reversion account is not an artifact of how we constructed v_{rev}. Put together, these results support a *gravitational interpretation of safety reversion*: benign adaptation contains a reproducible component toward an earlier, more permissive state, and suppressing that component limits safety degradation.

More concretely, our contributions are fourfold:

- **Independent operationalization and local interpretation.** We make the reversion hypothesis measurable with a heterogeneous family of independently trained helpful-only proxies, and show in the `Llama-3.1-8B` and `Qwen2.5-7B` settings that both the canonical and family directions carry a local request-engagement vocabulary signature on a disjoint set of harmful prompts.

- **Pre-specified geometric evidence.** Across reference sources, tasks, and seeds, benign fine-tuning contains the predicted reversion component, and every trajectory evaluated after the analysis was frozen exceeds a matched empirical null.

- **Behavioral relevance and directional control.** Reversion alignment tracks harmfulness along matched trajectories; an optimizer-matched block reduces safety degradation and a stabilized push amplifies it, with the same mean ordering across three model families from 3B to 14B.

- **Proof-of-concept safety-rehearsal triangulation.** A safety-rehearsal loss built without access to the direction supplies the predicted opposing component in Llama and Qwen, and reduces both cumulative reversion and unsafe rates.

We position these contributions relative to prior work on safety reversion and activation-space interventions in Section A.

2 THE GRAVITATIONAL INTERPRETATION

Harmless fine-tuning can increase the rate at which a safety-aligned model produces unsafe responses. Figure 2 shows this motivating pattern for `Llama-3.1-8B-Instruct` and `Qwen2.5-7B-Instruct` before we introduce a geometric interpretation and its testable predictions. We ask whether this behavioral drift has a reproducible geometric component. Alongside task-specific motion induced by the downstream objective, we hypothesize that the resulting trajectory contains a shared component along a reversion direction toward a more permissive state preceding safety alignment. To make this hypothesis measurable, we represent checkpoints through their residual-stream activations on a fixed probe family and use independently constructed helpful-only models as operational proxies for that earlier state. This construction defines the reversion direction from the safety-aligned starting model toward any helpful-only reference, against which subsequent benign adaptation can be decomposed.

We next formalize this construction, distinguish it from a within-model refusal direction, and derive its geometric and intervention predictions. Section 3.2.1 then describes the prospective evaluation protocol and its matched null.

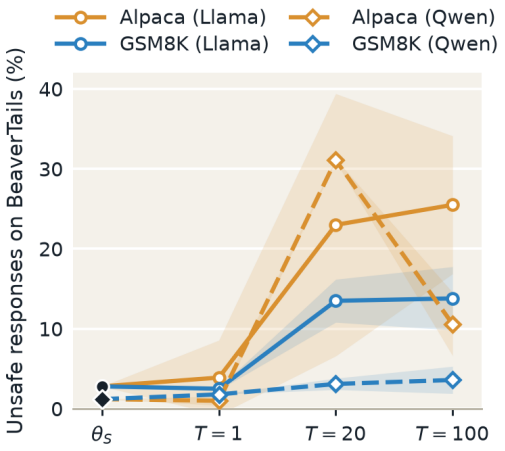


Figure 2: Early unsafe drift under ordinary benign full fine-tuning of `Llama-3.1-8B-Instruct` (solid) and `Qwen2.5-7B-Instruct` (dashed): BeaverTails completions judged unsafe by Llama Guard 3 ($n = 500$), mean $\pm$ std over two seeds. Llama LoRA runs and exact values are in Table 8.

2.1 MODELS STANDARD LIFECYCLE

Every checkpoint we study descends from the same pretrained base model θ_P. Safety alignment (*e.g.* RLHF against harmful requests) turns θ_P into the safety-aligned checkpoint θ_S that we start from; benign post-training adaptation on some downstream task then turns θ_S into a downstream checkpoint θ_T after T optimization steps.

Our hypothesis concerns a state that this lifecycle passes over. Evidence for basin-like loss landscapes suggests that pretraining establishes a broad basic-capability basin, while later alignment or specialization occupies model-dependent regions that are often narrower (Chen et al., 2026). Within that broad basin, before safety alignment, we posit a meta-state that is as permissive as the base model but already capable as a chat assistant, and θ_P itself does not qualify, since it is not yet a chatbot. Since neither this state nor the basin's global geometry is directly observable, we construct an operational proxy: a helpful-only reference θ_H branches from θ_P through a short helpful-only fine-tuning stage, keeping θ_P's lack of safety filtering while acquiring the chat-style instruction following that θ_S has and θ_P lacks. Helpful-only names this construction rather than the opposite of safety; data and filtering details are in Section B.2.

A broad basin contains many such states rather than a single one, so we construct several helpful-only references independently across data sources and seeds, each providing a local point from which a reversion direction relative to θ_S can be measured (Section 2.2). The resulting directions are closely aligned across sources and seeds (Table 3), so the family behaves as one coherent direction, summarized by its normalized *family centroid* v_{fam} (Section B.3). This coherence makes fixing one *canonical* reference per model–scale setting, and hence one canonical θ_H and v_{rev}, a practical rather than privileged choice: the interventions use it, and a matched block along v_{fam} reproduces its average effect (Section B.8).

2.2 BACKGROUND: ACTIVATION-SPACE DIRECTIONS

All directional measurements in the paper are defined in activation space. A *probe family* $P = \{x_i\}_{i=1}^{|P|}$ is a collection of prompts used to expose the behavior being studied. For checkpoint θ, let

$h_\theta(x_i, \ell)$ denote the residual-stream activation at layer ℓ and at the last non-padding token of prompt x_i. We summarize a checkpoint on that probe family by

$$\mu_\theta(P, \ell) = \frac{1}{|P|} \sum_{i=1}^{|P|} h_\theta(x_i, \ell). \tag{1}$$

The geometric objects below are displacements between these mean activations.

For a downstream checkpoint θ_T (Section 2.1), the representational shift from θ_S is

$$\Delta_T(P, \ell) = \mu_{\theta_T}(P, \ell) - \mu_{\theta_S}(P, \ell). \tag{2}$$

Given a helpful-only reference checkpoint θ_H, we define

$$v_{\text{rev}}(P, \ell) = \mu_{\theta_H}(P, \ell) - \mu_{\theta_S}(P, \ell). \tag{3}$$

Notably, we subtract activations only between architecture-matched checkpoints at the same layer and hidden coordinates. Each v_{rev} is therefore specific to its checkpoint pair and is recomputed rather than transferred across models or scales. We omit P and ℓ when they are clear from context.

Activation differences are also used to identify refusal features (Arditi et al., 2024). That construction compares one model across two prompt classes, for example

$$r_\ell = \mu_\theta(P_{\text{harmful}}, \ell) - \mu_\theta(P_{\text{harmless}}, \ell), \tag{4}$$

whereas Equation (3) compares two checkpoints on the same prompts. In the safety setting, harmful probes expose the contrast between the safety-aligned checkpoint and the helpful-only reference. We call v_{rev} a *reversion direction*; this geometric definition alone does not specify which output behavior the direction promotes. We test that separately in Section 3.3.1, where *request-engagement signature* names the measured local shift toward task-engaging rather than refusal-like opening tokens.

2.3 A PREFERRED COMPONENT OF LATER OPTIMIZATION

We now turn the hypothesis into falsifiable structure: a motivating premise sets the interpretive frame, and two predictions state what should be observed if the interpretation is correct.

Motivating premise — Training-scale asymmetry. Following Section 2.1, we read a helpful-only reference as a local proxy for an earlier permissive part of the broad pretraining regime, and θ_S as a later, narrower safety-aligned displacement from it. Distributional overlap offers one possible mechanism for reversion: helpful-only references already fit broad benign instruction-following data, so a later benign objective can produce genuine task learning while partially restoring earlier behavior from a safety-aligned starting point. The public base and instruction-tuned checkpoints are not a controlled before-and-after lineage, so this reading remains an interpretation; the testable claims below concern only the checkpoint-defined direction.

Geometric prediction — Subsequent fine-tuning contains a preferred reversion component. Consider a downstream run r (one task and one seed) saved at checkpoints $0 = T_0 < T_1 < \cdots < T_K$, and let

$$u_{r,k} = \mu_{\theta_{r,T_k}} - \mu_{\theta_{r,T_{k-1}}} \tag{5}$$

be the activation-space step between consecutive checkpoints. Projecting each step onto a unit direction $\hat{v}$ splits it into a coefficient $\alpha_{r,k} = \langle u_{r,k}, \hat{v} \rangle$ along $\hat{v}$ and an orthogonal residual $\epsilon_{r,k} = u_{r,k} - \alpha_{r,k}\hat{v}$, the two components sketched conceptually in Figure 1; this is an activation-space projection, not a parameter-space gradient decomposition. This split exists for every direction, so by itself it says nothing about v_{rev}. The prediction is that the direction fixed in advance from helpful-only references, the canonical v_{rev} or the family centroid v_{fam}, captures more of the observed motion than a matched null would. We measure this with the path-normalized score

$$A_r(v) = \frac{\langle \Delta_{r,T_K}, \hat{v} \rangle}{\sum_{k=1}^{K} \|u_{r,k}\|_2}, \tag{6}$$

the net displacement along $\hat{v}$ as a fraction of the path traveled through the saved checkpoints, which keeps runs that move by different amounts comparable. The prediction fails if these scores are not

consistently positive or do not exceed their matched sign-flip nulls across reference sources, downstream conditions, and seeds. It is a prediction of drift rather than collapse: reversion contributes a positive projection, while the residual $\epsilon_{r,k}$ collects the larger task-specific changes together with any other motion, and is therefore broader than a pure task gradient; the model need not return to a reference checkpoint for its trajectory to exhibit a systematic return bias.

Intervention prediction — Directional interventions should modulate outcomes. If v_{rev} captures a behaviorally relevant component, training objectives constructed to penalize or encourage motion along it should change downstream safety behavior.

Section 3 tests these predictions in order, exploring the geometric significance of our reversion direction, its correlation with the emergence of harmful behavior, and how it can be used to causally steer the harmlessness of the final post-trained checkpoint, and whether a standard safety-rehearsal objective independently opposes it.

3 Experiments

This section tests the empirical consequences of the interpretation in Section 2. We first ask whether post-alignment fine-tuning acquires a preferred component along v_{rev}, then whether that component co-evolves with unsafe drift in ordinary baseline runs and what it locally promotes in the next-token vocabulary. We next intervene with training objectives constructed from the measured direction and test whether they change the downstream behavioral outcome. Finally, we ask whether a safety-rehearsal loss independently supplies the opposing component predicted by the interpretation.

3.1 Setup

We study the Llama, Qwen, and Gemma families from 3B to 14B. Each model–scale setting is named after its base model (*e.g.* `Llama-3.1-8B`); within a setting, θ_S is the corresponding instruction-tuned release (`-Instruct`, or `-IT` for Gemma), and helpful-only references are trained from the base model. Benign downstream trajectories fine-tune θ_S on Alpaca, GSM8K, or structured JSON generation, with no harmful response targets in their task losses. Activation-space quantities are measured on 128 harmful AdvBench prompts (Zou et al., 2023) at a preselected late layer L^* per model (Section 2.2), and unsafe behavior is the fraction of completions on a fixed, disjoint 500-prompt BeaverTails subset (Ji et al., 2024) judged unsafe by Llama Guard 3. Per-experiment coverage, training recipes, reference construction, and extraction details are in Sections B.1 and B.2.

3.2 Directional emergence of the reversion component

The first question is directional rather than behavioral: does subsequent fine-tuning acquire a growing component along a reference-defined reversion direction, rather than drifting arbitrarily in representation space? For `Llama-3.1-8B`, the reference family contains nine helpful-only models, three training seeds on each of UltraChat (Ding et al., 2023), Dolly-15k (Conover et al., 2023), and WildChat (Zhao et al., 2024); their directions have mean pairwise cosine 0.863 and an effective dimension of 1.29 out of nine (Table 3).

3.2.1 Cross-fitted preferred-direction test

We test Equation (6) with whole trajectories as experimental units, building each direction on four of five prompt folds and comparing held-out scores against a matched null that sign-flips centered prompt-level reference differences, preserving their variability while removing the mean direction (Section B.3). After a four-trajectory pilot, we froze the reference family and decision rule and trained five further trajectories across Alpaca, GSM8K, and JSON: all five exceed their null $p95$, jointly accounting for 29.5% of the saved-checkpoint path against a null $p95$ of 6.8%, and every direction and trajectory in the pooled nine-run suite exceeds its null. The component emerges early, with centroid cosine rising from 0.436 ± 0.063 at $T = 1$ to 0.662 ± 0.112 at $T = 100$ (Figure 3), and scores well below one place it on top of larger task- and seed-specific motion rather than a collapse onto one axis. In a matched `Llama-3.2-3B` comparison, the helpful-only direction also aligns more strongly than the raw base-model direction (Table 4).

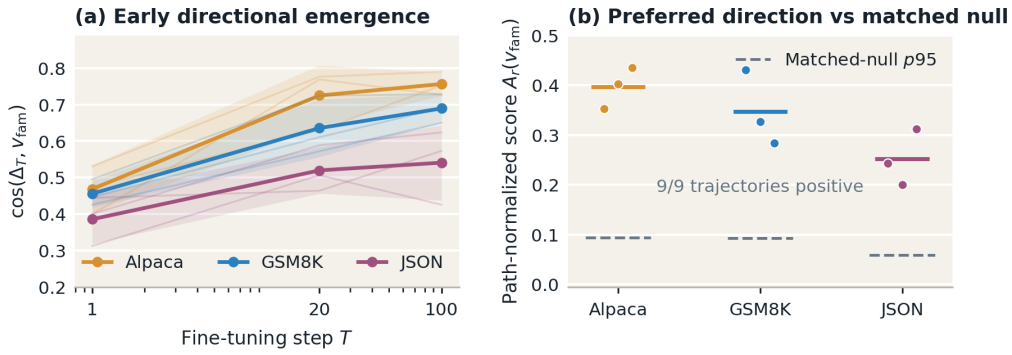


Figure 3: Preferred reversion across nine `Llama-3.1-8B` baseline trajectories. **(a)** Cosine to the cross-fitted family centroid; faint lines are seeds, bands mean ± std. **(b)** Path-normalized centroid score per run; solid segments are task means, dashed segments task-matched null $p95$.

Table 1: Cross-fitted family-centroid cosine $C_{\mathrm{CF}}(r, T)$ and BeaverTails unsafe rate (%) during ordinary full fine-tuning of `Llama-3.1-8B-Instruct`.

		$T = 1$		$T = 20$		$T = 100$	
Task	**Seed**	**Cos.**	**Unsafe**	**Cos.**	**Unsafe**	**Cos.**	**Unsafe**
Alpaca	0	0.3989	0.6	0.7690	34.6	0.7271	19.4
Alpaca	1	0.4753	7.2	0.6297	11.4	0.7524	31.6
GSM8K	0	0.4951	2.6	0.7243	15.4	0.7285	16.6
GSM8K	1	0.4461	2.4	0.6107	11.6	0.6900	11.0

3.3 Behavioral coupling of trajectory alignment and unsafe drift

We next ask, observationally, whether the component that optimization naturally follows co-evolves with unsafe behavior over the same early interval. On ordinary Alpaca and GSM8K full fine-tuning of `Llama-3.1-8B-Instruct`, we pair the cross-fitted family-centroid cosine $C_{\mathrm{CF}}(r, T)$, computed with the held-out-fold construction above, with the BeaverTails unsafe rate at the same checkpoints (Table 1). In all four runs, both quantities are higher at $T = 100$ than at $T = 1$, and mean unsafe rate rises at every sampled checkpoint; this paired $4/4$ endpoint result is the primary summary, while the pooled Spearman $r = 0.958$ (Pearson 0.860; Spearman 0.974 under Qwen3Guard) describes the time courses rather than 12 independent replications. Baselines, LoRA runs, and second-judge results are in Sections B.5 and B.11.

3.3.1 Local vocabulary interpretation and probe dependence

To ask what this coordinate locally promotes, we measure at the aligned `Llama-3.1-8B-Instruct` checkpoint the exact derivative of a fixed next-token contrast between affirmative, task-engaging and refusal-like openings. On 64 harmful AdvBench prompts disjoint from direction construction, $\hat{v}_{\mathrm{rev}}$ and $\hat{v}_{\mathrm{fam}}$ yield means of 0.462 and 0.434 log-mass-contrast units per unit residual-stream displacement; both are positive on all 64 prompts, exceed every one of $10{,}000$ matched prompt-variability nulls ($p_{\mathrm{MC}} = 1/10{,}001$), and survive equal token weighting. Benign probes recover related but non-identical directions (canonical cosine 0.811; harmful–benign family-centroid cosine 0.635). We therefore describe v_{rev} as carrying a local *request-engagement signature*, not as a unique compliance or refusal axis; full estimands, nulls, and checks are in Section B.4.

3.4 Intervention relevance of the reversion component

Observation alone leaves open that v_{rev} is only a correlate, so we intervene during benign Alpaca fine-tuning from θ_S with an auxiliary objective built from the fixed, model-specific direction. The *block* is a one-sided barrier against positive projection, optimizer-matched to ordinary fine-tuning,

and gives the primary comparison; the *push* reinforces the direction through a separately stabilized bounded angular objective and provides complementary evidence. We evaluate baseline, block, and push at $T = 100$ in six model–scale settings (Table 2); objectives and hyperparameters are in Section B.6.

For `Llama-3.1-8B`, blocking motion along v_{rev} flips the geometric drift itself (Figure 4), moving from $\cos(\Delta_{100}, v_{\text{rev}}) = 0.666 \pm 0.032$ in the baseline to -0.208 ± 0.022, and reduces BeaverTails harmfulness from $24.6\% \pm 8.3\%$ to $5.7\% \pm 1.9\%$. This is a reduction of safety degradation rather than full recovery: harmfulness remains above the 2.8% aligned-model anchor, although the block closes most of the gap opened by ordinary benign fine-tuning. Across the same matched runs, block shows no aggregate additional degradation relative to ordinary fine-tuning in held-out Alpaca assistant-token NLL (1.389 ± 0.010 against 1.404 ± 0.031) or IFEval ($52.1\% \pm 8.7\%$ against $40.0\% \pm 12.7\%$; Tables 9 and 13). A matched block along the family centroid v_{fam} reproduces this mean effect ($\cos = -0.190 \pm 0.047$, harmfulness $7.5\% \pm 9.0\%$; Section B.8). Push moves in the opposite direction, raising alignment to 0.742 ± 0.013 and harmfulness to $48.8\% \pm 4.3\%$; the same mean ordering persists under Qwen3Guard and on HarmBench (Mazeika et al., 2024) (Table 15).

Random-axis control. In a separate paired suite, four of five calibrated random-axis blocks also lower mean harmfulness under both judges, by 15.6 points on average under Llama Guard (9.3 over all five axes, against 9.0 for the canonical block run alongside them), while remaining positively aligned with v_{rev} (mean cosine 0.467), so the safety benefit is shared. Across these ten runs, harmfulness tracks the remaining alignment with v_{rev} (descriptive Spearman $\rho = 0.64$), and what sets v_{rev} apart is that its natural alignment and intervention sign are fixed ex ante, whereas which random axes help is known only after evaluation (Section B.6 and fig. 5).

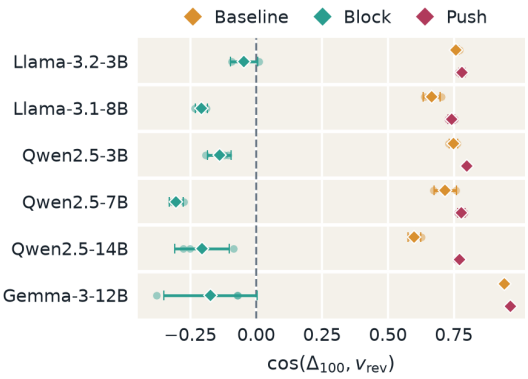


Figure 4: Cosine to v_{rev} at $T = 100$ in all six settings; dots are runs, diamonds mean $\pm$ std over three seeds.

Robustness across late layers and scales. Without layer-specific coefficient retuning, block lowers mean harmfulness at every tested late layer and improves safety in 14 of 15 matched layer–seed comparisons (Table 14). Across all six model–scale settings, each with its own aligned checkpoint, canonical reference, and local v_{rev}, mean alignment and harmfulness are lowest under block, intermediate under ordinary fine-tuning, and highest under push (Figure 4 and table 2); this is persistence of the ordering across scales, not monotonic effect-size growth. Across all 54 runs at $T = 100$, alignment and harmfulness are also rank-correlated (Spearman 0.73, $p = 3.5 \times 10^{-10}$), with within-setting values between 0.80 and 1.00. Coefficients were frozen before multi-seed expansion (Table 12); all per-model metrics are in Table 13.

3.5 Safety rehearsal supplies an implicit counter-component

Whereas the block uses v_{rev} explicitly, we ask as an external check whether a safety-retaining objective with no access to the direction nevertheless opposes it. We study one common form, safety rehearsal: a joint objective $\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{retain}}$ in which $\mathcal{L}_{\text{retain}}$ rehearses the aligned model's own safe responses to harmful requests, leaving weight- or output-space regularization toward θ_S to future tests. If rehearsal counteracts the reversion component, matched one-step updates from a fixed state should satisfy $\langle u_{\text{retain}}, \hat{v}_{\text{rev}} \rangle < 0$, and joint training should reduce the cumulative projection $p(T) = \langle \Delta_T, \hat{v}_{\text{rev}} \rangle$, *i.e.* $p_{\text{joint}}(T) < p_{\text{task}}(T)$. Neither inequality is guaranteed by fitting safe textual targets, since the sign is measured on a separate harmful probe set against a direction constructed independently from helpful-only references; we test both from Llama and Qwen states with paired task-only and joint trajectories (Section B.10).

Table 2: Unsafe BeaverTails responses (%) at $T = 100$; mean$_{\pm\text{std}}$ over three seeds, θ_S a single evaluation. ΔNLL and ΔIFEval are block minus baseline in held-out Alpaca NLL and IFEval prompt-level strict accuracy (points). Full metrics in Table 13.

Setting	θ_S	Baseline	Block	Push	ΔNLL	ΔIFEval
`Llama-3.2-3B`	5.2	$18.3_{\pm 2.8}$	$1.7_{\pm 1.1}$	$30.5_{\pm 2.7}$	$+0.017$	$+11.1$
`Llama-3.1-8B`	2.8	$24.6_{\pm 8.3}$	$5.7_{\pm 1.9}$	$48.8_{\pm 4.3}$	-0.015	$+12.1$
`Qwen2.5-3B`	1.4	$15.7_{\pm 2.9}$	$6.4_{\pm 1.1}$	$26.3_{\pm 3.8}$	0.000	$+0.3$
`Qwen2.5-7B`	1.2	$17.1_{\pm 4.0}$	$3.6_{\pm 0.5}$	$42.8_{\pm 4.8}$	-0.001	$+1.2$
`Qwen2.5-14B`	0.8	$9.7_{\pm 5.8}$	$3.2_{\pm 0.9}$	$14.5_{\pm 1.2}$	$+0.003$	$+1.2$
`Gemma-3-12B`	2.2	$17.2_{\pm 4.2}$	$9.9_{\pm 3.8}$	$33.8_{\pm 12.9}$	-0.020	-2.4

Every retaining-only update has negative projection onto v_{rev} across all 12 Llama and all four Qwen states, while task-update alignment is positive on average (mean local cosine -0.359 and -0.404 for the retaining updates against $+0.112$ and $+0.303$ for the task updates, for Llama and Qwen respectively); the shared structure is specifically the retaining update's component along $-v_{\text{rev}}$, not a whole-vector negative of the task update (whole-vector cosine 0.09 for Llama and -0.37 for Qwen). During joint training, retaining reduces cumulative projection at $T = 100$ by $30.9\% \pm 1.8\%$ for Llama and $30.3\% \pm 14.1\%$ for Qwen. On BeaverTails, unsafe rate falls from $20.7\% \pm 6.9\%$ to $0.4\% \pm 0.3\%$ for Llama and from $18.5\% \pm 2.4\%$ to $1.9\% \pm 0.6\%$ for Qwen; on HarmBench it falls from $63.0\% \pm 0.0\%$ to $29.0\% \pm 9.9\%$ for Llama and from $57.7\% \pm 2.5\%$ to $40.7\% \pm 7.4\%$ for Qwen (Figure 6); complete behavioral and utility results are in Section B.10.

4 Discussion

What the results establish. Reversion directions built independently from helpful-only references predict a shared component of later benign adaptation beyond a matched prompt-variability null; this component co-evolves with harmfulness along ordinary trajectories; and objectives built from it change downstream safety in the predicted order, with block mitigating and push amplifying degradation across the tested families and scales. Safety rehearsal, which never sees the direction, independently supplies the opposing component.

Interpretation and scope. These findings motivate a local gravitational interpretation: task optimization produces task-specific and other changes, but repeatedly contains an additional component toward the helpful-only proxy for an earlier, more permissive state. The proxy is operational rather than an observed ancestor, and the path-normalized scores remain far below one: trajectories are far from one-dimensional, with reversion as one component alongside dominant task-specific motion, and the evidence concerns local geometry around each aligned checkpoint rather than a global basin. Each control bounds the claim in a specific way. The sign-flip null supports a shared directional component but leaves checkpoint-level mean structure and distributional overlap outside its scope; the vocabulary readout gives that component a request-engagement signature, with both prompt-general and harmful-prompt-dependent structure; and the baseline–block contrast attaches the causal claim to an objective built from v_{rev} rather than to a unique coordinate mediator, since optimization can also alter other coordinates, as the imperfect correspondence between projection and harmfulness under block and rehearsal also shows. Finally, several calibrated random blocks also reduce harmfulness; what distinguishes v_{rev} is that it is fixed ex ante and naturally followed by ordinary benign adaptation.

Implications and limitations. The block is a diagnostic rather than a proposed deployment safeguard, while rehearsal shows that ordinary text-level safety training can supply a related counter-component. The source-diverse prospective test uses `Llama-3.1-8B`, while model-specific directions and interventions span three model families from 3B to 14B; these evaluations broaden, but do not universalize, the scope. Accordingly, "gravitational" denotes a reproducible local tendency in the studied benign regimes, not a literal physical law, global potential, or universal scaling relation; extended qualifications are given in Section C.

ETHICS STATEMENT

This work studies why safety behavior degrades under benign fine-tuning, with the aim of understanding and limiting that degradation. All experiments use publicly released open-weight models and public benchmarks (AdvBench, BeaverTails, HarmBench). The push objective, which increases unsafe responses, serves only as a diagnostic of whether the reversion direction is behaviorally relevant, rather than as a proposed method, and fine-tuning directly on harmful data is a simpler and already documented way to remove safety behavior (Qi et al., 2024). Unsafe completions are scored by automatic safety classifiers; the qualitative examples in the appendix are truncated, and specific drug names are blurred. The block objective is likewise a diagnostic: its reductions in unsafe rate are evidence of the direction's behavioral relevance, and certifying it as a deployment safeguard would require separate robustness evaluation.

REPRODUCIBILITY STATEMENT

The appendix documents every experiment at the level needed to reproduce it: software versions, optimizer settings, and tokenization (Section B.1); helpful-only reference data, filtering, training, and activation extraction (Section B.2); the cross-fitted preferred-direction test, its matched null, Monte Carlo seeds, and the decision rule fixed before the prospective trajectories were trained (Section B.3); the vocabulary readout estimands and nulls (Section B.4); the intervention objectives and per-model hyperparameters (Section B.6); and the safety-rehearsal data, including how prompts and safe targets were selected (Section B.10). All models, datasets, and safety classifiers are public, and anonymized code for all experiments is provided as supplementary material.

AI USE STATEMENT

We used large language models to assist with drafting and revising sections of this manuscript, retrieving and summarizing related work, and running the experiments reported in the paper, including launching and monitoring experimental jobs and collecting their outputs.

REFERENCES

Mohammed Talha Alam, Nada Saadi, Fahad Shamshad, Nils Lukas, Karthik Nandakumar, Fahkri Karray, and Samuele Poppi. SPQR: A multi-dimensional benchmark for safety alignment under benign model adaptation. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2511.19558.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11717.

Thong Bach, Dung Nguyen, Thao Minh Le, and Truyen Tran. Continual safety alignment via gradient-based sample selection. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://arxiv.org/abs/2604.17215.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. URL https://arxiv.org/abs/2204.05862.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.17424.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://arxiv.org/abs/2402.11746.

Huanran Chen, Zeming Wei, Yao Huang, Yichi Zhang, Yinpeng Dong, and Jun Zhu. Unveiling the basin-like loss landscape in large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2505.17646.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. Databricks, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023. URL https://arxiv.org/abs/2305.14233.

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2505.06843.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. URL https://arxiv.org/abs/2004.10964.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.01099.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. We should chart an atlas of all the world's models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2503.10633.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.16833.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2402.01109.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.01586.

Yuhan Huang, Huanran Chen, and Yinpeng Dong. Alignment dynamics in llm fine-tuning. *arXiv preprint arXiv:2605.18309*, 2026. URL https://arxiv.org/abs/2605.18309.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language models resist alignment: Evidence from data compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23411–23432, Vienna, Austria, July 2025. Association

for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1141. URL https://aclanthology.org/2025.acl-long.1141/.

Jiaming Ji et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*, 2024.

Sanghyun Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Safety alignment backfires: Preventing the re-emergence of suppressed concepts in fine-tuned text-to-image diffusion models. *arXiv preprint arXiv:2412.00357*, 2024. URL https://arxiv.org/abs/2412.00357.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.02155.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.17374.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. Towards understanding the fragility of multilingual LLMs against fine-tuning attacks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2358–2372, 2025. doi: 10.18653/v1/2025.findings-naacl.126. URL https://aclanthology.org/2025.findings-naacl.126/.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.03693.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.05946.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025. URL https://arxiv.org/abs/2506.11618.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.00761.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.05162.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2109.01652.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024. URL https://arxiv.org/abs/2405.01470.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.