# Continual Safety Alignment via Gradient-Based Sample Selection

**Thong Bach**[1*] **Dung Nguyen**[1] **Thao Minh Le**[2] **Truyen Tran**[1]

[1] Applied Artificial Intelligence Initiative (A2I2), Deakin University [2] Pennsylvania State University

## Abstract

Large language models require continuous adaptation to new tasks while preserving safety alignment. However, fine-tuning on even benign data often compromises safety behaviors, including refusal of harmful requests, truthfulness, and commonsense reasoning. We investigate which training samples cause alignment drift through a data-centric lens. Our empirical analysis shows samples contribute unequally: high-gradient samples cause greater safety degradation and drive models toward pretrained distributions, while moderate-gradient samples enable task learning with minimal alignment loss. We propose gradient-based sample selection that filters high-gradient samples during fine-tuning. Across multiple model families on continual domain tasks, our method substantially improves alignment preservation while maintaining competitive task performance, without requiring curated safe data or architectural modifications. Our method is robust across selection ratios, task orderings, and diverse attack benchmarks.

## 1 Introduction

Large language models deployed in real-world applications require continuous adaptation to new domains, tasks, and evolving requirements. While initial alignment through reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), direct preference optimization (DPO) (Rafailov et al., 2024), and constitutional AI (Bai et al., 2022) establishes safety properties, subsequent fine-tuning often compromises these carefully cultivated behaviors (Qi et al., 2024a; Yang et al., 2023).

This vulnerability presents a fundamental challenge for LLM deployment. Organizations need to customize models for specific use cases, incorporate new knowledge, and adapt to changing requirements, yet each fine-tuning step risks degrading the alignment properties that make these models safe to deploy. Even fine-tuning on benign, non-malicious datasets can unintentionally weaken safety mechanisms (Qi et al., 2024a; He et al., 2024), suggesting that alignment degradation is not merely a consequence of adversarial data but a structural property of fine-tuning itself. While the data content is typically benign, the parameter updates they induce can be destructive to the alignment priors.

Continual learning research has made significant progress on retaining task performance through parameter regularization (Kirkpatrick et al., 2017) and experience replay (Rolnick et al., 2019). However, these methods focus on preventing catastrophic forgetting of learned tasks rather than preserving alignment properties. We address a distinct problem:

**The Continual Safety Alignment Problem**

How can we continuously adapt LLMs to new tasks while preserving alignment properties (safety, truthfulness, helpfulness) without requiring curated safe data at each adaptation step?

Rather than constraining model architectures or requiring curated safe datasets at each adaptation step, we investigate this problem through a data-centric lens: *Which training samples cause alignment drift, and can we simply avoid them?*

Recent work reveals two critical properties of aligned models that inform our approach. First, (Ji et al., 2024) shows that LLMs exhibit *elasticity*: a tendency to revert toward pretrained distributions during fine-tuning because the massive pretraining corpora exerts stronger influence than smaller alignment datasets. This reversion inherently degrades alignment since these pre-training corpora usually lack safety constraints. Second, (Peng et al., 2024) discovers that aligned models occupy a "safety

* t.bach@deakin.edu.au



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 18870–18887

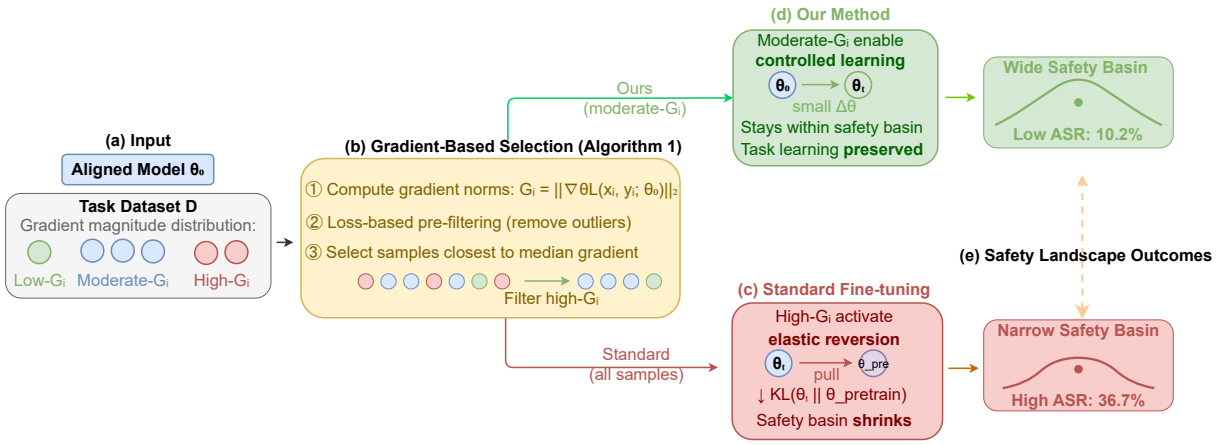


Figure 1: **Overview of gradient-based sample selection for continual safety alignment.** (a) Aligned model $\theta_0$ and task dataset with varying gradient magnitudes. (b) Our method filters high-$G_i$ samples (red) and selects moderate-$G_i$ samples (blue). (c) Standard fine-tuning activates elastic reversion, pulling models toward pretrained distributions. (d) Our method enables controlled updates preserving alignment. (e) Resulting safety landscapes: narrow basin (high ASR) vs. wide basin (low ASR). High-gradient samples may represent alignment tension points where training tends to reverse safety modifications.

basin" in parameter space with sharp boundaries where safety collapses abruptly.

While these frameworks explain *why* alignment is fragile, they do not predict *which samples* cause drift or *how* to maintain alignment during task adaptation. We hypothesize that training samples activate elastic reversion unequally: samples where aligned predictions diverge substantially from task targets (high gradients) may reverse alignment modifications, while moderate-gradient samples enable learning with minimal drift.

We validate this hypothesis through systematic experiments and propose a practical data-centric solution. Our contributions are:

**Empirical finding.** We provide empirical evidence that per-sample gradient magnitude predicts safety drift in continual fine-tuning. Through KL-divergence analysis, we show high-gradient samples shift models toward pretrained distributions (Table 3), and demonstrate that standard remedies such as gradient clipping are insufficient (Section E.4), establishing that the issue is sample-specific rather than purely magnitude-based.

**Mechanistic analysis.** We characterize high-gradient samples as format mismatches—short-answer tasks (classification, closed QA) where the aligned model's verbose output distribution diverges from terse targets—connecting our gradient-based findings with independent observations from representation-based data selection (Hsiung et al., 2025) and benign data auditing (He et al., 2024).

**Practical recipe.** We propose gradient-based sample selection with a tunable safety-task trade-off, validated across three model families, multiple task sequences and orderings, and diverse safety benchmarks including AdvBench and HarmBench. Our method requires no curated safe data or architectural modifications.

## 2 Background

### 2.1 Measuring Alignment via Safety Basins

To study alignment dynamics quantitatively, we adopt the safety basin framework from (Peng et al., 2024). This framework conceptualizes alignment as occupying a region in parameter space rather than a binary property.

Given aligned parameters $\theta_{\text{align}}$, the safety landscape is defined by perturbing along direction $\hat{d}$: $f(\alpha) = S(\theta_{\text{align}} + \alpha\hat{d})$ where $S(\cdot)$ measures attack success rate. This defines a *safety basin* $\mathcal{B} = \{\theta : S(\theta) \leq S_{\text{threshold}}, \theta \text{ connected to } \theta_{\text{align}}\}$, the connected region in parameter space where safety properties hold.

A critical empirical finding is that safety basins have *sharp boundaries*: safety exhibits step-function collapse when crossing the boundary, with minimal graceful degradation. This geometry makes large parameter updates particularly dangerous. A single large step can push models from safe to unsafe, while many small updates might stay within the basin.

The VISAGE score quantifies basin volume by averaging safety margin across random perturba-

tion directions:

$$\text{VISAGE} = \mathbb{E}_{\alpha \sim U(-a,a)}[S_{\max} - S(\alpha)] \quad \text{s.t. } S < S_{\max}, \tag{1}$$

where $S_{\max} = 100\%$ represents complete safety failure. Higher VISAGE indicates larger safety basins and more robust alignment. We define alignment drift at step $t$ as: $\Delta_{\text{align}}(t) = \text{VISAGE}(\theta_0) - \text{VISAGE}(\theta_t)$. In our experiments, we compute VISAGE using $N = 100$ random perturbation directions, with perturbation range $a$ calibrated per model. The safety margin is averaged across all directions, providing robust measurement beyond any single perturbation slice. See Appendix B for extended discussion.

### 2.2 Alignment Fragility and Elasticity

Alignment degradation during fine-tuning reflects structural properties of the learning process rather than just adversarial data. Previous work (Qi et al., 2024a) shows that as few as 10 examples can compromise safety, while (He et al., 2024) demonstrates that even benign datasets weaken safety mechanisms.

The elasticity framework (Ji et al., 2024) provides theoretical grounding for this phenomenon: language models resist alignment modifications and rebound toward pretrained behavior under perturbation. The elastic force is proportional to dataset size: $F_{\text{elastic}} \propto |\mathcal{D}_i| \cdot \Delta D_{\text{KL}}(p_\theta \| p_{\mathcal{D}_i})$. Since pretrain corpora ($|\mathcal{D}_p|$) vastly exceed alignment datasets ($|\mathcal{D}_a|$), the pretrained distribution exerts orders of magnitude stronger "pull" on model behavior.

This asymmetry predicts two phenomena: *resistance* (pretrained models resist initial alignment) and *rebound* (aligned models revert toward pretrained behavior under fine-tuning). Our hypothesis extends this framework to the sample level: training samples where aligned predictions diverge substantially from task targets, indicated by high gradient magnitudes, may specifically activate this elastic reversion force, accelerating drift toward pretrained distributions and out of safety basins.

### 2.3 Problem Formulation

We formalize the continual safety alignment problem, distinguishing it from standard continual learning.

**Setting.** Consider an aligned model $\theta_0$ that must learn $T$ tasks sequentially from datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$. Each dataset $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ contains input-output pairs for task $t$.

**Standard Continual Learning.** Traditional continual learning minimizes task loss while preventing forgetting of previous tasks:

$$\theta_t = \arg\min_\theta \mathcal{L}_t(\theta; \mathcal{D}_t) + \lambda \cdot \text{Reg}(\theta; \theta_{t-1}) \tag{2}$$

where $\text{Reg}(\cdot)$ penalizes deviation from previous parameters (e.g., EWC (Kirkpatrick et al., 2017)) or replays stored examples (Rolnick et al., 2019).

**Continual Safety Alignment.** We impose an additional constraint, alignment preservation:

$$\theta_t = \arg\min_\theta \mathcal{L}_t(\theta; \mathcal{D}_t) \quad \text{s.t.} \quad \theta_t \in \mathcal{B} \tag{3}$$

where $\mathcal{B}$ is the safety basin. The model must remain within the basin throughout training, not just at convergence. Since pretrained models lie outside the safety basin and elasticity pulls parameters toward pretrained configurations, fine-tuning must resist this force to preserve alignment.

**Alignment Drift Metric.** We quantify alignment degradation using VISAGE (Eq. 1). Define alignment drift at step $t$ as:

$$\Delta_{\text{align}}(t) = \text{VISAGE}(\theta_0) - \text{VISAGE}(\theta_t) \tag{4}$$

The continual safety alignment objective becomes:

$$\min_{\theta_1, \ldots, \theta_T} \sum_{t=1}^{T} \mathcal{L}_t(\theta_t; \mathcal{D}_t) \quad \text{s.t.} \quad \Delta_{\text{align}}(t) \leq \delta \quad \forall t \tag{5}$$

where $\delta$ is a tolerance threshold for acceptable alignment drift.

**Challenges.** The constraint in Eq. 5 is difficult to enforce directly: **1. Measurement cost**: Computing VISAGE requires evaluating safety across multiple perturbation directions, which is expensive during training. **2. Non-differentiability**: VISAGE is not differentiable with respect to $\theta$, precluding gradient-based constraint enforcement. **3. Sample-level opacity**: The constraint operates at the model level, but training operates at the sample level, so it is unclear which samples contribute to drift.

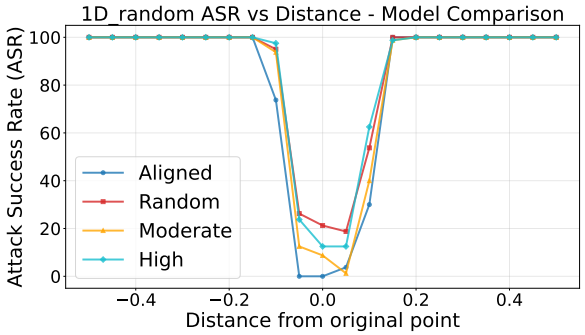


Figure 2: Safety landscape visualization for Qwen-2.5-7B-Instruct after fine-tuning on Dolly. Each curve shows the attack success rate under parameter perturbations. High-$G_i$ selection (cyan) creates a narrow basin where small perturbations cause safety collapse. Moderate-$G_i$ (orange) maintains a wide basin comparable to the original model (blue), demonstrating robust alignment preservation.

**Toward a data-centric solution.** The third challenge, sample-level opacity, motivates our approach. Rather than enforcing model-level constraints (expensive and non-differentiable), we ask: *which training samples accelerate alignment drift?* If samples contribute unequally, we can filter those causing disproportionate drift. This shifts the problem from constrained optimization to sample selection.

## 3 Analysis: Which Samples Cause Drift?

Following the data-centric perspective introduced in Section 2, we investigate: *do all samples contribute equally to alignment drift, or can sample-level properties predict drift risk?*

**Hypothesis.** We hypothesize that per-sample gradient magnitude indicates drift risk. High-gradient samples occur where aligned predictions diverge substantially from task targets, precisely where alignment training modified behavior away from pretrained tendencies. Training on these samples may reverse those modifications, activating elastic reversion.

**Experimental design.** We fine-tune LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024) on Dolly (Conover et al., 2023) (15K benign instruction-following examples). We compare three selection strategies, each using 3,000 samples (20%): (1) **Random** (baseline), (2) **High-$G_i$** (top 20% by gradient norm $G_i = \|\nabla_\theta \mathcal{L}(x_i, y_i; \theta_0)\|_2$), and (3) **Moderate-$G_i$** (20% closest to median gradient norm). All use identical hyperparameters (lr=$2 \times 10^{-5}$, 3 epochs, AdamW). We evaluate VISAGE (Peng et al., 2024), attack success rate (ASR) on AdvBench (Zou et al., 2023), and task performance. See Appendix C for full details.

| Model | Selection | ASR↓ | TruthfulQA | Task Avg |
|---|---|---|---|---|
| Qwen2.5-7B | Low-$G_i$ | **8.4** | **50.2** | 60.1 |
| | Moderate-$G_i$ | 10.2 | 42.5 | **60.9** |
| LLaMA-3.1-8B | Low-$G_i$ | **13.3** | **45.5** | 44.5 |
| | Moderate-$G_i$ | 18.3 | 41.5 | **46.4** |
| Qwen3-4B | Low-$G_i$ | **3.0** | **48.6** | 56.3 |
| | Moderate-$G_i$ | 6.0 | 42.8 | **58.1** |

Table 1: Low-$G_i$ vs. Moderate-$G_i$ selection. Low-$G_i$ provides better safety preservation but trades 0.8–1.9 points of task performance, revealing a Pareto tradeoff along the gradient spectrum.

| Model | Selection | VISAGE | Retain | ASR↓ |
|---|---|---|---|---|
| Qwen-2.5-7B | Aligned | 78.5 | - | 2.1 |
| | High-$G_i$ | 48.8 | 62.2% | 18.4 |
| | Random | 57.0 | 72.6% | 12.7 |
| | Moderate-$G_i$ | **65.5** | **83.4%** | **5.8** |
| LLaMA-3.1-8B | Aligned | 67.7 | - | 3.2 |
| | High-$G_i$ | 48.9 | 72.2% | 15.1 |
| | Random | 52.8 | 78.0% | 9.8 |
| | Moderate-$G_i$ | **59.3** | **87.6%** | **4.9** |

Table 2: Alignment preservation across selection strategies. High-gradient selection causes the largest degradation (62-72% retention); moderate-gradient selection preserves 83-88% with the lowest ASR. Results shown for two model families.

**Why not low-gradient samples?** To characterize the full gradient-safety spectrum, we evaluate Low-$G_i$ selection (bottom 20% by gradient norm) as a complete baseline across all three model families (Table 1).

Low-$G_i$ provides better safety preservation across all models but consistently trades 0.8–1.9 points of task performance. This reveals a clean Pareto tradeoff: Low-$G_i$ for best safety, Moderate-$G_i$ for best task performance, High-$G_i$ worst on both dimensions. This strengthens our core finding that gradient magnitude monotonically predicts alignment drift. We recommend Moderate-$G_i$ as a practical default; safety-critical applications may prefer Low-$G_i$.

### 3.1 Results Support the Hypothesis

**High-gradient samples show greater alignment drift** Table 2 shows high-$G_i$ selection retains only 62-72% of original alignment and increases ASR by 5-9×, while moderate-$G_i$ selection preserves 83-88% with only 1.5-2× ASR increase.

| Model | Selection | $KL_{pretrain}$ | $KL_{aligned}$ |
|---|---|---|---|
| Qwen-2.5-7B | High-$G_i$ | 0.14 | 0.25 |
| | Random | 0.15 | 0.23 |
| | Moderate-$G_i$ | **0.18** | **0.17** |
| LLaMA-3.1-8B | High-$G_i$ | 0.51 | 0.14 |
| | Random | 0.55 | 0.11 |
| | Moderate-$G_i$ | **0.68** | **0.06** |

Table 3: KL-divergence to pretrained and aligned distributions of different sample selection methods. High-$G_i$ samples are associated with shifts toward pretrained distributions (lower $KL_{pretrain}$) and away from aligned distributions (higher $KL_{aligned}$).

| Model | Parameter | HIGH | MOD | r | p |
|---|---|---|---|---|---|
| Qwen2.5-7B | Last_V | 0.119 | 0.104 | 0.41 | $< 10^{-3}$ |
| | Last_O | 0.276 | 0.244 | 0.39 | $< 10^{-3}$ |
| | Middle | −0.004 | −0.004 | 0.06 | 0.38 |
| LLaMA-3.1-8B | Last_MLP | 0.104 | 0.102 | 0.18 | $< 0.01$ |
| | Last_V | −0.029 | −0.033 | 0.33 | $< 10^{-3}$ |
| | Middle | −0.020 | −0.024 | 0.03 | 0.72 |

Table 4: TopK-Cosine ($k$=1000) between gradients and reversion direction. HIGH and MOD denote the top 20% and middle 20% by gradient norm. Directional alignment appears in final-layer parameters with no effect in middle layers.

Figure 2 visualizes this: high-$G_i$ dramatically narrows the safety basin, while moderate-$G_i$ preserves basin width comparable to the original model.

**Evidence for elastic reversion mechanism.** Table 3 provides evidence consistent with the elastic reversion hypothesis. Training on High-$G_i$ samples correlates with movement toward pretrained distributions (lower $KL_{pretrain}$) and *away from* aligned distributions (higher $KL_{aligned}$). This pattern is consistent with high-gradient samples activating the elastic reversion force described by (Ji et al., 2024): they represent alignment tension points where training reverses safety modifications.

### 3.2 Gradient Direction Analysis: Preliminary Investigation

To explore whether high-gradient samples have gradients aligned with the reversion direction $\mathbf{r} = \theta_{pretrain} - \theta_{aligned}$, we analyze gradient directions across parameter subsets. Direct cosine similarity in billion-dimensional parameter spaces yields near-zero values due to concentration of measure. We address this by computing TopK-Cosine: cosine similarity restricted to the $k$ dimensions where alignment training induced the largest parameter changes (see Appendix D for methodology details).

Table 4 presents results for Qwen2.5-7B-Instruct and LLaMA-3.1-8B-Instruct. High-gradient samples exhibit higher directional alignment with the reversion direction compared to moderate-gradient samples in final-layer parameters, though the specific components vary by architecture: V/O projections in Qwen2.5 (TopK-Cosine 0.119 vs 0.104 for V, $r = 0.41$) and MLP layers in LLaMA (0.104 vs 0.102, $r = 0.18$). Besides, middle layers show no directional effect in either model ($|r| < 0.07$, $p > 0.3$), confirming that the signal localizes to alignment-critical parameters in the final transformer layer.

These findings provide preliminary support for the elastic reversion hypothesis: *high-gradient samples may produce gradients that partially reverse final-layer modifications learned during alignment.* However, the modest correlation strengths ($r = 0.18$–$0.41$) and architectural variation (attention outputs in Qwen2.5 vs. MLP in LLaMA) suggest this relationship is complex. We note that our selection method relies only on gradient magnitude, not direction, and is effective regardless of whether the directional hypothesis fully holds. We emphasize the gradient clipping comparison (Section E.4) and empirical stability as the primary support for why accumulation of small gradient steps does not undermine our method.

### 3.3 Implications for Method Design

These findings validate our data-centric approach: *filter high-gradient samples during fine-tuning*. Moderate-gradient samples provide sufficient learning signal for task adaptation while avoiding alignment-reversing effects. This motivates our gradient-based selection method (Section 4), which requires no curated safe data, only gradient computation on the task dataset itself.

## 4 Method: Gradient-Based Sample Selection

Our analysis reveals that moderate-gradient samples enable task learning with minimal alignment drift. We operationalize this insight through a batch selection algorithm that filters high-gradient samples during fine-tuning.

### 4.1 Algorithm

Algorithm 1 operates in three stages: (1) loss-based pre-filtering removes extreme samples (very low

| $\rho$ | ASR↓ | TruthfulQA | HellaSwag | Task Avg |
|---|---|---|---|---|
| 0.1 | 2.7 | 43.3 | 69.7 | 59.3 |
| **0.2** | **6.0** | **42.8** | **70.1** | **58.1** |
| 0.4 | 5.5 | 43.9 | 69.3 | 57.1 |
| 0.6 | 6.6 | 42.2 | 69.5 | 57.2 |

Table 5: Sensitivity to selection ratio $\rho$ on Qwen3-4B. ASR remains consistently low across $\rho \in [0.1, 0.4]$ (2.7–6.0%), all substantially better than baseline (16.6%) and random (11.8%).

loss = memorized; very high loss = outliers), retaining $\sim$68% of candidates, (2) gradient computation on filtered candidates, reducing computational cost by not computing gradients for all candidates, and (3) median-based selection chooses samples closest to median gradient norm $\mu_G$, avoiding both high-gradient (causing alignment drift) and low-gradient samples (providing minimal learning signal).

**Algorithm 1** Gradient-Based Sample Selection

**Require:** Batch $\mathcal{B}$, model $\theta$, selection ratio $\rho = 0.2$

**Ensure:** Selected samples $\mathcal{S}$

1: Compute losses: $L_i = \mathcal{L}(x_i, y_i; \theta)$ for all $(x_i, y_i) \in \mathcal{B}$
2: Filter: $\mathcal{C} \leftarrow \{(x_i, y_i) : L_i \in [\mu_L - \sigma_L, \mu_L + \sigma_L]\}$
3: Compute gradient norms: $G_i = \|\nabla_\theta \mathcal{L}(x_i, y_i; \theta)\|_2$ for $(x_i, y_i) \in \mathcal{C}$
4: $\mu_G \leftarrow \text{median}(\{G_i\})$
5: Select $\lfloor \rho|\mathcal{B}| \rfloor$ samples closest to $\mu_G$ as $\mathcal{S}$
6: **return** $\mathcal{S}$

**Key design choices.** We use median (not mean) for robustness against heavy-tailed gradient distributions. Selection ratio $\rho \in [0.15, 0.25]$ balances quality vs. cost; we use $\rho = 0.2$. Pre-filtering reduces gradient computation by $\sim$32%.

### 4.2 Sensitivity to Selection Ratio $\rho$

We conduct a systematic sensitivity analysis on $\rho \in \{0.1, 0.2, 0.4, 0.6\}$ using Qwen3-4B across the full 4-task continual learning pipeline (Table 5).

Results are robust across $\rho \in [0.1, 0.4]$: ASR remains consistently low (2.7–6.0%), all substantially better than baseline (16.6%) and random sampling (11.8%). Smaller $\rho$ (stricter filtering) provides slightly better safety (2.7% at $\rho = 0.1$) at marginal task performance cost (59.3% vs 58.1%). Larger $\rho$ (less filtering) converges toward random sampling behavior. We recommend $\rho = 0.2$ as a default, but practitioners can tune based on safety-task priorities.

## 5 Experiments

We evaluate our gradient-based sample selection method on continual safety alignment across multiple model families and diverse task sequences. Our experiments demonstrate that filtering high-gradient samples better preserves both task performance and safety alignment throughout sequential fine-tuning.

### 5.1 Experimental Setup

**Models and tasks.** We evaluate three instruction-tuned models: Qwen2.5-7B-Instruct (Yang et al., 2024), LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen3-4B-Instruct (Yang et al., 2025) (abbreviated as Qwen2.5, LLaMA-3.1, and Qwen3). Following realistic deployment scenarios, we fine-tune sequentially on four domains: Dolly (Conover et al., 2023) (general instruction-following), GSM8K (Cobbe et al., 2021) (mathematical reasoning), MedMCQA (Pal et al., 2022) (medical QA), and Squad_v2 (Rajpurkar et al., 2018) (reading comprehension). Dolly is used first to reduce excessive refusal behavior in RLHF-aligned models.

**Training configuration.** All models use LoRA (Hu et al., 2022) (rank 32, $\alpha = 64$), AdamW optimizer, learning rate $10^{-4}$, batch size 32, and one epoch per task. For Moderate-$G_i$, $\rho = 0.2$. Results averaged over three seeds.

**Baselines.** (1) **Baseline**: standard fine-tuning; (2) **Random**: random sampling; (3) **KL**: KL-divergence regularization against aligned model; (4) **O-LoRA** (Wang et al., 2023): orthogonal subspace learning; (5) **EWC** (Kirkpatrick et al., 2017): Elastic Weight Consolidation, a Fisher-based continual learning method; (6) **Grad. Clip**: gradient norm clipping with clip value 0.5 (best performing among $\{0.1, 0.5, 1.0\}$; see Section E.4).

**Evaluation.** Task performance via lm-evaluation-harness (Gao et al., 2024). Alignment via: (1) ASR on AdvBench (Zou et al., 2023), (2) ASR on HarmBench (Mazeika et al., 2024), encompassing direct requests, contextual attacks, and optimization-based jailbreaks, (3) TruthfulQA (Lin et al., 2022), (4) commonsense reasoning (ARC-C, BoolQ, HellaSwag, Winogrande). We report checkpoint-averaged values (mean $\pm$ std over three seeds) unless otherwise noted; the high ASR vari-

| Model | Method | ASR↓ | TruthfulQA | ARC-C | BoolQ | HellaSwag | Winogrande |
|---|---|---|---|---|---|---|---|
| Qwen2.5 | Baseline | 36.7 ± 13.6 | 38.2 ± 1.2 | 58.0 ± 1.5 | **86.4** ± 1.0 | 79.2 ± 0.5 | **72.3** ± 0.3 |
| | Random | 31.1 ± 14.3 | 38.4 ± 1.1 | 58.1 ± 0.6 | 85.9 ± 2.0 | 79.4 ± 0.3 | **72.3** ± 0.5 |
| | KL | 33.5 ± 13.4 | 37.8 ± 1.3 | 58.3 ± 1.7 | 86.2 ± 1.4 | 79.1 ± 0.4 | **72.3** ± 0.4 |
| | O-LoRA | 16.5 ± 19.7 | **42.9** ± 1.0 | 56.6 ± 0.9 | 86.1 ± 1.2 | 79.3 ± 0.4 | 71.8 ± 0.6 |
| | EWC | 17.4 ± 6.7 | 38.1 ± 0.3 | 57.7 ± 0.4 | 86.8 ± 0.2 | 79.2 ± 0.1 | 71.7 ± 0.3 |
| | Grad. Clip | 31.2 ± 13.8 | 38.2 ± 1.1 | 57.9 ± 1.4 | 86.3 ± 1.1 | 79.3 ± 0.4 | 72.1 ± 0.4 |
| | Moderate-$G_i$ | **10.2** ± 7.1 | 42.5 ± 1.3 | **59.6** ± 1.7 | 86.3 ± 1.5 | **79.9** ± 0.1 | 71.7 ± 0.9 |
| LLaMA-3.1 | Baseline | 44.2 ± 22.5 | 37.8 ± 0.7 | 56.3 ± 1.3 | 84.8 ± 1.0 | 77.9 ± 0.5 | 73.8 ± 0.5 |
| | Random | 31.9 ± 23.3 | 38.6 ± 1.7 | **56.9** ± 1.5 | 84.8 ± 0.6 | 78.0 ± 0.3 | 73.6 ± 0.5 |
| | KL | 43.6 ± 21.6 | 38.6 ± 0.8 | 56.4 ± 1.5 | **84.9** ± 1.0 | 78.0 ± 0.6 | **74.0** ± 0.5 |
| | O-LoRA | 23.3 ± 28.0 | 40.0 ± 1.0 | 56.4 ± 1.2 | 84.7 ± 0.6 | **78.3** ± 0.6 | **74.0** ± 0.7 |
| | EWC | 40.2 ± 11.8 | 38.0 ± 0.3 | 56.0 ± 0.7 | 84.9 ± 0.2 | 78.0 ± 0.1 | 74.0 ± 0.2 |
| | Moderate-$G_i$ | **18.3** ± 17.3 | **41.5** ± 2.6 | 56.0 ± 1.5 | 84.8 ± 0.6 | 77.9 ± 0.7 | 73.7 ± 0.5 |
| Qwen3 | Baseline | 16.6 ± 7.7 | 39.7 ± 0.7 | 59.8 ± 1.5 | **86.2** ± 0.5 | **71.0** ± 1.2 | 68.7 ± 0.4 |
| | Random | 11.8 ± 5.8 | 40.1 ± 1.6 | **60.2** ± 1.1 | 85.6 ± 1.2 | 70.6 ± 0.8 | 68.7 ± 0.7 |
| | KL | 17.8 ± 7.0 | 39.7 ± 0.8 | 59.5 ± 1.2 | 86.0 ± 0.4 | **71.0** ± 1.2 | **69.0** ± 0.7 |
| | O-LoRA | 6.8 ± 6.5 | 42.3 ± 1.6 | 59.5 ± 1.3 | 85.2 ± 0.8 | 70.7 ± 1.1 | 68.7 ± 0.7 |
| | EWC | 14.1 ± 3.1 | 40.5 ± 0.2 | 59.3 ± 0.3 | 85.7 ± 0.1 | 71.1 ± 0.1 | 68.9 ± 0.3 |
| | Moderate-$G_i$ | **6.0** ± 5.4 | **42.8** ± 1.6 | 59.2 ± 1.3 | 84.6 ± 1.4 | 70.1 ± 0.8 | 68.5 ± 0.8 |

Table 6: Alignment preservation metrics, checkpoint-averaged (mean ± std over three seeds). Safety via attack success rate (ASR, lower better), truthfulness via TruthfulQA, and general capabilities via ARC-Challenge, BoolQ, HellaSwag, and Winogrande.

ances in some rows reflect natural variation across the four training stages rather than instability within a single stage.

### 5.2 Alignment Preservation

We first evaluate safety and general capabilities averaged across all training checkpoints and three seeds. Table 6 presents attack success rates (ASR), truthfulness (TruthfulQA), and general capabilities (ARC-C, BoolQ, HellaSwag, Winogrande).

**Safety preservation.** Moderate-$G_i$ achieves substantially lower attack success rates throughout continual learning. On Qwen2.5, our method achieves 10.2% ASR versus 36.7% (Baseline), 31.1% (Random), 33.5% (KL), and 16.5% (O-LoRA)—representing 3.6× reduction over Baseline. EWC achieves 17.4% ASR but notably *worsens* safety on LLaMA-3.1 (40.2% vs. 44.2% baseline), demonstrating that Fisher-based regularization, designed to protect task-critical parameters, does not address alignment preservation. Gradient clipping provides only marginal improvement (31.2% on Qwen2.5), confirming that the issue is sample-specific rather than purely magnitude-based (see Section E.4). On Qwen3, Moderate-$G_i$ (6.0%) matches O-LoRA (6.8%), both substantially outperforming Baseline (16.6%) and Random (11.8%). On LLaMA-3.1, Moderate-$G_i$ (18.3%) improves over Baseline (44.2%) and KL (43.6%). KL regularization provides minimal benefit despite explicitly constraining distribution drift, while O-LoRA performs well but requires architectural modifications.

These results validate our hypothesis: high-gradient samples reverse safety training during fine-tuning, and filtering them preserves alignment guardrails.

**Broader safety evaluation.** To assess generalization beyond AdvBench, we evaluate on HarmBench across the full continual learning pipeline on LLaMA-3.1-8B (Table 9). Our method achieves 5.6× lower ASR than baseline on HarmBench, with consistent improvements over all baselines, including EWC, demonstrating that safety gains generalize to more diverse attack vectors.

**Cross-architecture variation.** Safety improvements vary across families: Qwen2.5 shows 3.6× ASR reduction, Qwen3 shows 2.8×, while LLaMA-3.1 shows 2.4× with no task performance gains over baseline. This variation likely reflects differences in alignment procedures or architectural factors.

**Truthfulness and general capabilities.** Moderate-$G_i$ maintains factual accuracy (42.5-42.8% on TruthfulQA, matching or exceeding baselines) and general capabilities (competitive on ARC-C, BoolQ, HellaSwag, Winogrande, with

| Model | Method | After GSM8K | After MedMCQA | After Squad | Avg |
|---|---|---|---|---|---|
| Qwen2.5 | Baseline | 58.2 ± 0.2 | 57.8 ± 0.8 | 61.3 ± 0.9 | 59.1 |
| | Random | 57.7 ± 0.7 | 56.8 ± 4.1 | 62.2 ± 2.6 | 58.9 |
| | KL | 58.1 ± 0.2 | **63.1** ± 0.3 | 61.7 ± 2.1 | **61.0** |
| | O-LoRA | 54.7 ± 0.6 | 59.2 ± 0.5 | 53.0 ± 0.8 | 55.7 |
| | EWC | 57.8 ± 0.2 | 57.3 ± 1.2 | 63.3 ± 0.5 | 59.5 |
| | **Moderate-$G_i$** | **59.0** ± 0.5 | 60.6 ± 1.0 | **63.0** ± 2.7 | 60.9 |
| LLaMA-3.1 | Baseline | 44.4 ± 0.5 | 43.8 ± 1.0 | 51.5 ± 1.0 | 46.5 |
| | Random | 44.4 ± 1.3 | 44.2 ± 0.9 | 51.9 ± 1.9 | 46.8 |
| | KL | 44.5 ± 0.7 | 43.9 ± 0.6 | 50.7 ± 2.1 | 46.4 |
| | O-LoRA | 43.6 ± 0.9 | **45.7** ± 0.6 | **54.5** ± 0.9 | **47.9** |
| | EWC | 43.8 ± 0.7 | 44.3 ± 0.5 | 50.5 ± 1.2 | 46.2 |
| | **Moderate-$G_i$** | **45.2** ± 1.9 | 44.4 ± 0.5 | 49.7 ± 1.9 | 46.4 |
| Qwen3 | Baseline | 60.7 ± 0.2 | 51.2 ± 0.6 | 50.6 ± 1.0 | 54.2 |
| | Random | 60.6 ± 0.9 | 54.1 ± 3.0 | 50.2 ± 3.9 | 55.0 |
| | KL | 59.8 ± 0.9 | 52.4 ± 1.8 | 52.2 ± 2.4 | 54.8 |
| | O-LoRA | **60.8** ± 0.3 | **59.1** ± 1.0 | 56.6 ± 0.4 | **58.8** |
| | EWC | 59.2 ± 0.4 | 53.1 ± 0.7 | 52.6 ± 1.6 | 55.0 |
| | **Moderate-$G_i$** | 58.2 ± 0.4 | 57.6 ± 3.3 | **58.5** ± 3.2 | 58.1 |

Table 7: Checkpoint-averaged performance across all evaluation tasks at each training stage. Values represent mean ± standard deviation over three seeds. Moderate-$G_i$ maintains the highest overall performance throughout continual learning.

| Model | Method | Avg Perf. | BWT↑ | FM↓ | Max Drop↓ |
|---|---|---|---|---|---|
| Qwen2.5 | Baseline | 59.1 | −1.7 | 1.7 | 11.8 |
| | Random | 58.9 | +0.4 | −0.4 | 10.6 |
| | KL | 61.0 | −0.9 | 2.8 | 5.0 |
| | O-LoRA | 55.7 | −8.8 | 11.4 | 21.4 |
| | Moderate-$G_i$ | **60.9** | −1.8 | 1.8 | **2.7** |
| LLaMA-3.1 | Baseline | 46.5 | +0.2 | −0.2 | 8.4 |
| | Random | 46.8 | +0.8 | −0.8 | 4.6 |
| | KL | 46.4 | −1.1 | 1.1 | 8.7 |
| | O-LoRA | **47.9** | **+6.4** | **−4.5** | – |
| | Moderate-$G_i$ | 46.4 | −1.7 | 1.7 | 5.4 |
| Qwen3 | Baseline | 54.2 | −18.5 | 18.5 | 32.5 |
| | Random | 55.0 | −19.8 | 19.8 | 23.2 |
| | KL | 54.8 | −15.5 | 15.5 | 26.0 |
| | O-LoRA | 58.8 | −12.4 | 12.4 | 15.3 |
| | **Moderate-$G_i$** | **58.1** | **−4.3** | **4.3** | **5.6** |

Table 8: Continual learning performance with forgetting metrics. Avg Perf.: checkpoint-averaged accuracy. BWT: Backward Transfer (higher is better). FM: Forgetting Measure (lower is better). Max Drop: worst single-step performance drop. Moderate-$G_i$ achieves 14.2% BWT improvement and 5.8× reduction in max drop on Qwen3.

differences typically within 1-2 points). Sample selection preserves model competencies while improving safety. The complementary strengths of O-LoRA suggest that combining data-centric selection with architectural constraints may yield further improvements.

| Method | HarmBench ASR↓ |
|---|---|
| Baseline | 27.8 (3.7) |
| Random | 17.2 (5.6) |
| KL | 27.7 (1.0) |
| EWC | 31.0 (6.7) |
| O-LoRA | 10.0 (1.4) |
| Moderate-$G_i$ | **5.0** (3.4) |

Table 9: HarmBench ASR on LLaMA-3.1-8B across the full continual learning pipeline. Moderate-$G_i$ achieves 5.6× reduction over baseline, generalizing to diverse attack types.

### 5.3 Continual Learning Performance

Table 7 presents checkpoint-averaged performance, computed as the mean accuracy across all three evaluation tasks (GSM8K, MedMCQA, Squad_v2) at each training stage. This metric captures both the model's competence on the current task and its retention of previous capabilities, which is the central challenge in continual learning.

Our gradient-based selection consistently outperforms baselines in maintaining balanced performance across tasks. On Qwen2.5, Moderate-$G_i$ achieves 60.9% average performance, a 2.0 percentage point improvement over Random (58.9%) and 1.8 points over Baseline (59.08%). The advantage grows more pronounced on Qwen3, where our method achieves 58.1% versus 55.0% for Ran-

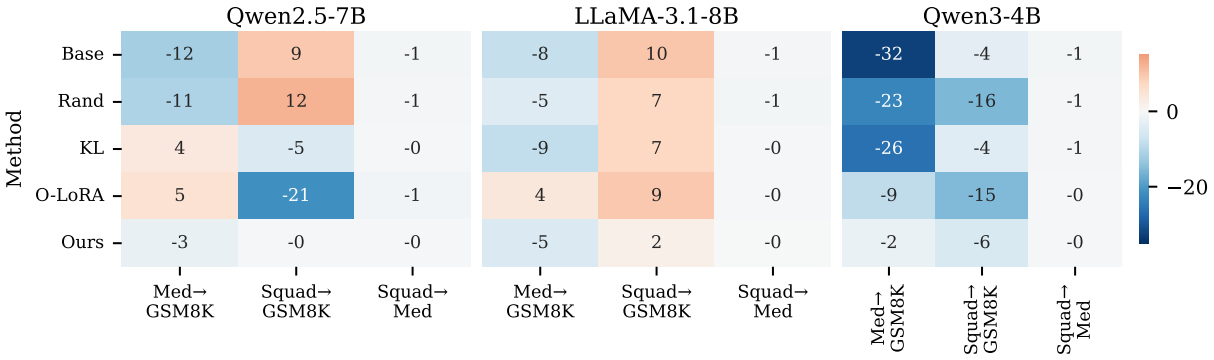


Figure 3: Task interference matrix. Negative values (blue) indicate interference; positive (red) indicates transfer. Moderate-$G_i$ consistently minimizes cross-task interference.

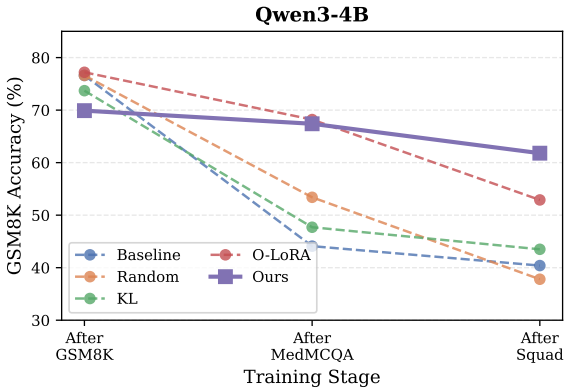


Figure 4: GSM8K trajectory on Qwen3. Moderate-$G_i$ maintains gradual degradation while baselines exhibit catastrophic drops.

dom and 54.2% for Baseline, representing a 3.9 percentage point gain.

### 5.4 Catastrophic Forgetting Analysis

We analyze catastrophic forgetting using standard continual learning metrics from (Lopez-Paz and Ranzato, 2017): *Backward Transfer* (BWT), measuring how learning new tasks affects previous task performance, and *Forgetting Measure* (FM), quantifying the gap between peak and final accuracy.

Table 8 extends our performance results with these metrics. On Qwen3, Moderate-$G_i$ achieves BWT of $-4.3\%$ compared to $-18.5\%$ for Baseline—a **14.2 percentage point improvement**. The forgetting measure confirms this: 4.3% versus 18.5%, a $4.3\times$ reduction. These gains are largest on Qwen3, where baselines suffer severe forgetting.

**Task Interference Analysis.** Figure 3 reveals *why* methods differ: MedMCQA training causes $-32.5\%$ interference with GSM8K for Baseline on Qwen3, while Moderate-$G_i$ experiences only $-2.5\%$—a **30 percentage point reduction**. Total task interference drops from $-37.1\%$ to $-8.5\%$ ($4.4\times$ reduction), explaining why gradient-based selection preserves prior knowledge.

**Forgetting Trajectories.** Figure 4 tracks GSM8K accuracy on Qwen3. Moderate-$G_i$ retains 61.8% at the final checkpoint (8.1% drop from peak), while Baseline and Random collapse to 44.1% and 53.4% after MedMCQA training. The "Max Drop" column in Table 8 quantifies this: Moderate-$G_i$ limits worst-case drops to 5.6% versus 32.5% for Baseline ($5.8\times$ reduction), preventing the catastrophic single-step forgetting that risks crossing safety basin boundaries.

## 6 Conclusion

We presented an empirical investigation of which training samples cause alignment drift during continual fine-tuning. Our findings show that high-gradient samples accelerate reversion toward pre-trained distributions, while moderate-gradient samples enable task learning with minimal alignment loss. Our gradient-based sample selection filters high-gradient samples during fine-tuning, achieving strong alignment preservation on Qwen models and modest improvements on LLaMA-3.1, with consistent gains across task orderings, diverse safety benchmarks (AdvBench, HarmBench), and selection ratios. We demonstrate that standard remedies such as gradient clipping and EWC are insufficient, establishing that the issue is sample-specific. While the 51% computational overhead may limit applicability for very large models, our findings demonstrate that sample selection provides a practical, architecture-agnostic approach to continual safety alignment.

## References


Meta AI. 2024. Llama guard 3: A safeguard for human-ai conversations. https://github.com/meta-llama/PurpleLlama.

Thong Bach, Dung Nguyen, Thao Minh Le, and Truyen Tran. 2026a. Rethinking deep alignment through the lens of incomplete safety learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30005–30012.

Thong Bach, Thanh Nguyen-Tang, Dung Nguyen, Thao Minh Le, and Truyen Tran. 2026b. Curvature-aware safety restoration in LLMs fine-tuning. *Transactions on Machine Learning Research*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. 2025. Benign samples matter! fine-tuning on outlier benign samples severely breaks safety. In *Forty-second International Conference on Machine Learning*.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What's in your "safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*.

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. 2025. Why llm safety guardrails collapse after fine-tuning: A similarity analysis between alignment and fine-tuning datasets. *arXiv preprint arXiv:2506.05346*.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Safe LoRA: The silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. 2024a. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024b. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024c. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. *arXiv preprint arXiv:2405.18641*.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024d. Vaccine: Perturbation-aware alignment for large language models. *arXiv preprint arXiv:2402.01109*.

Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Juntao Dai, Yunhuai Liu, and Yaodong Yang. 2024. Language models resist alignment: Evidence from data compression. *arXiv preprint arXiv:2406.06144*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Yan-Shuo Liang and Wu-Jun Li. 2025. Gated integration of low-rank adaptation for continual learning of language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30.

Ning Lu, Shengcai Liu, Jiahao Wu, Weiyu Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. 2025. Safe delta: Consistently preserving safety when fine-tuning llms on diverse datasets. *arXiv preprint arXiv:2505.12038*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Jishnu Mukhoti, Yarin Gal, Philip HS Torr, and Puneet K Dokania. 2023. Fine-tuning can cripple your foundation model; preserving features may be the solution. *arXiv preprint arXiv:2308.13320*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

ShengYun Peng, Pin-Yu Chen, Jianfeng Chi, Seongmin Lee, and Duen Horng Chau. 2025. Shape it up! restoring llm safety during finetuning. *arXiv preprint arXiv:2505.17196*.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. 2024. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024a. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024b. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. 2024. Representation noising effectively prevents harmful fine-tuning on llms. *arXiv preprint arXiv:2405.14577*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, and 1 others. 2024. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. *arXiv preprint arXiv:2501.13453*.

Andy Zhou, Bo Li, and Haohan Wang. 2024. Robust prompt optimization for defending language models against jailbreaking attacks. *Advances in Neural Information Processing Systems*, 37:40184–40211.

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

## A Related Work

**Alignment Fragility and Theoretical Foundations.** Safety alignment is surprisingly fragile. Fine-tuning on as few as 10 adversarial examples can compromise safety, while even benign datasets like Alpaca and Dolly inadvertently degrade alignment (Qi et al., 2024b; Wei et al., 2023; Bach et al., 2026b). These vulnerabilities extend to various attack vectors and parameter-efficient methods (Yang et al., 2023; Lermen et al., 2023; Guan et al., 2025). The elasticity framework provides theoretical grounding: language models exhibit an inherent tendency to revert toward pretrained distributions, with elastic force proportional to dataset size (Ji et al., 2024). This predicts both resistance to initial alignment and rebound toward pretrained behavior under fine-tuning. A complementary geometric perspective conceptualizes alignment through "safety basins" in parameter space, where safety exhibits sharp, step-function collapse at basin boundaries (Peng et al., 2024). We adopt the VISAGE metric and extend this framework to understand sample-level contributions to alignment drift.

**Continual Learning for LLMs.** Classical continual learning addresses catastrophic forgetting through parameter regularization (Kirkpatrick et al., 2017) and experience replay (Rolnick et al., 2019). LLM-specific studies reveal that forgetting intensifies with model scale (Luo et al., 2023) and that vertical continuity (general to specific) differs from horizontal continuity (across time and domains) (Shi et al., 2024). Parameter-efficient approaches like O-LoRA learn tasks in orthogonal subspaces to minimize interference (Wang et al., 2023). However, these methods focus on task performance rather than alignment preservation, which constitutes a distinct challenge (Zheng et al., 2025; Liang and Li, 2025).

**Defenses Against Safety Degradation.** Defenses vary by intervention stage. *Alignment-stage* methods immunize models before fine-tuning through perturbation-aware optimization (Huang et al., 2024d), mapping harmful representations to noise (Rosati et al., 2024; Mukhoti et al., 2023; Huang et al., 2024c; Lu et al., 2025), or meta-learning to simulate attacks (Tamirisa et al., 2024; Liu et al., 2023; Zhou et al., 2024; Huang et al., 2024b). *Fine-tuning-stage* methods constrain adaptation by separating optimization states with proximal regularization (Huang et al., 2024c) or incorporating safety examples during training. *Post-fine-tuning* methods (Peng et al., 2025; Bach et al., 2026a) recover safety by projecting weights onto safety-aligned subspaces (Hsu et al., 2024) or pruning harmful parameters (Huang et al., 2024a).

**Distinction from One-Shot Safe Fine-Tuning.** While the above defenses address important challenges, they primarily target single-step fine-tuning scenarios. Continual safety alignment differs in requiring: (1) robustness across sequential adaptations rather than one event, (2) operation without curated safe data at each step, and (3) accounting for cumulative drift across tasks. Our gradient-based selection addresses these issues through a data-centric approach that requires only gradient computation on task data. We compare against KL regularization and O-LoRA as baselines that operate in the continual setting without requiring safe data; a comprehensive comparison against all fine-tuning defenses would require adapting them to sequential multi-task settings.

## B Extended Background

This appendix provides an extended discussion of the theoretical foundations underlying our work: the safety basin framework and the elasticity phenomenon.

### B.1 Safety Basin Framework: Detailed Treatment

The safety basin framework, introduced by (Peng et al., 2024), provides a geometric perspective on LLM alignment that fundamentally changes how we understand alignment stability. Rather than treating alignment as a binary property or a single

metric, this framework reveals that aligned models occupy a connected region in parameter space with distinctive geometric properties.

**Safety Landscape Construction.** Given an aligned model with parameters $\theta_{\text{align}}$, the safety landscape is constructed by perturbing along a direction $\mathbf{d}$:

$$f(\alpha) = S\big(\theta_{\text{align}} + \alpha\hat{\mathbf{d}}\big) \tag{6}$$

where $S(\cdot)$ is a safety metric (e.g., attack success rate on adversarial benchmarks) and $\hat{\mathbf{d}}$ is a normalized perturbation direction. For 2D visualization, two orthogonal directions are used: $f(\alpha, \beta) = S(\theta_{\text{align}} + \alpha\hat{\mathbf{d}}_1 + \beta\hat{\mathbf{d}}_2)$.

**Key Empirical Properties.** (Peng et al., 2024) establish several universal properties across popular open-source LLMs including LLaMA-2, LLaMA-3, Mistral, and Vicuna. First, they observe a **flat interior**: random perturbations within the basin preserve safety, meaning the aligned model is not a sharp local optimum but sits within a stable region where small random changes to model weights do not immediately compromise safety. Second, there is a **sharp boundary** where safety exhibits step-function collapse when crossing the basin boundary with minimal graceful degradation. Models are essentially either safe or unsafe, which contrasts sharply with the capability landscape where performance peaks at the origin and gradually declines with perturbation. Third, examining **fine-tuning trajectories** reveals that fine-tuning on harmful data drags models out of the basin, while fine-tuning on mixed (harmful + safe) data can keep models within the basin, suggesting data composition affects trajectory direction. Finally, the framework reveals significant **system prompt sensitivity**: removing default system prompts or using roleplaying prompts can reduce VISAGE scores substantially.

**VISAGE Metric Details.** The VISAGE (Volumetric Index for Safety Alignment Guided by Explanation) score quantifies basin volume:

$$\text{VISAGE} = \mathbb{E}_{\alpha\sim U(-a,a)}[S_{\max} - S(\alpha)] \quad \text{s.t.} \quad S < S_{\max} \tag{7}$$

where $a$ is the perturbation range and $S_{\max}$ is the maximum safety violation score. In practice, this is computed by sampling $N$ random perturbation directions (we use $N = 100$), evaluating safety at multiple perturbation magnitudes for each direction, and computing the average safety margin across all evaluations. Higher VISAGE indicates a larger safety basin and more robust alignment. Empirically, VISAGE scores correlate with model resilience to fine-tuning attacks: models with higher initial VISAGE require more harmful data to break alignment.

**Contrast with Capability Landscape.** The safety basin geometry differs fundamentally from the capability landscape. Capability performance peaks at the trained parameters and degrades *gradually* with perturbation magnitude. In contrast, safety is *flat* within the basin and collapses *abruptly* at the boundary. This asymmetry has important implications: while capability loss under fine-tuning is typically gradual and recoverable, safety loss can be sudden and catastrophic.

### B.2 Elasticity Framework: Extended Discussion

The elasticity framework (Ji et al., 2024) provides theoretical grounding for understanding why alignment is fragile under fine-tuning, drawing on insights from data compression theory.

**Compression-Theoretic Foundation.** (Ji et al., 2024) model language model training through the lens of data compression. A language model $p_\theta$ trained on dataset $\mathcal{D}$ achieves compression rate $\gamma_{p_\theta}^{\mathcal{D}} = H(\mathcal{D})/H_{p_\theta}(\mathcal{D})$, where $H(\mathcal{D})$ is the entropy of the data and $H_{p_\theta}(\mathcal{D})$ is the cross-entropy under the model.

**Elastic Force Formulation.** The elastic force exerted by dataset $\mathcal{D}_i$ on model parameters is $F_{\text{elastic}} \propto |\mathcal{D}_i| \cdot \Delta D_{\text{KL}}(p_\theta \| p_{\mathcal{D}_i})$. This formulation reveals a critical asymmetry: since pretrain corpora ($|\mathcal{D}_p|$) vastly exceed alignment datasets ($|\mathcal{D}_a|$), the pretrained distribution exerts orders of magnitude stronger "pull" on model behavior.

**Resistance and Rebound.** The elasticity framework predicts two phenomena. **Resistance** occurs because pretrained models resist initial alignment due to data volume asymmetry; the alignment process must overcome the elastic force from the massive pretrain corpus. **Rebound** is the counterintuitive finding that more deeply aligned models revert *faster* to pretrained behavior under perturbation, because deeper alignment represents a larger deviation from the stable pretrained configuration, creating stronger restoring force.

**Experimental Validation.** (Ji et al., 2024) validate elasticity through several experiments. Training to reverse alignment (moving from aligned to unaligned) consistently shows lower training loss than forward alignment, confirming the asymmetric difficulty. Models trained with more positive (aligned) data initially perform better but deteriorate faster when fine-tuned with negative data. Larger models exhibit stronger rebound effects, with faster initial performance decline and slower subsequent decline.

**Connection to Our Work.** We extend the elasticity framework to the sample level. While (Ji et al., 2024) characterizes aggregate behavior under fine-tuning, we show that individual samples activate elastic reversion unequally. High-gradient samples, where the aligned model's predictions diverge substantially from fine-tuning targets, represent points of "alignment tension" that preferentially activate the reversion force.

# C Extended Experimental Setup

## C.1 Model Details

We conduct experiments on three model families representing different architectures and alignment procedures. **LLaMA 3.1 8B Instruct** (Grattafiori et al., 2024) is Meta's instruction-tuned model with safety alignment through RLHF, featuring 32 layers, 32 attention heads, and 4096 hidden dimension. **Qwen-2.5 7B Instruct** (Yang et al., 2024) is Alibaba's instruction-tuned model with multilingual capabilities, using 28 layers, 28 attention heads, and 3584 hidden dimension. **Qwen3-4B Instruct** (Yang et al., 2025) is a smaller variant with efficient architecture: 24 layers, 24 attention heads, and 2560 hidden dimension. All models have undergone alignment training and exhibit safety behaviors such as refusing harmful requests and providing helpful responses.

## C.2 Dataset Details

**Dolly.** The Databricks Dolly dataset (Conover et al., 2023) contains 15,000 instruction-following examples distributed across diverse categories: open QA (2,584 examples), closed QA (1,750), summarization (1,250), information extraction (1,500), creative writing (1,500), classification (2,000), and brainstorming (4,416). Dolly is a benign dataset containing no adversarially designed harmful content, making it suitable for studying unintentional alignment degradation.

**Continual Learning Tasks.** We use GSM8K (Cobbe et al., 2021) containing 8,500 grade-school math problems requiring multi-step reasoning, MedMCQA (Pal et al., 2022) with medical multiple-choice questions from Indian medical entrance exams, and SQuAD v2 (Rajpurkar et al., 2018) for reading comprehension, including unanswerable questions.

## C.3 Evaluation Details

**Attack Success Rate (ASR).** We evaluate safety using AdvBench (Zou et al., 2023) with 520 harmful queries. To assess whether model responses are harmful, we employ Llama-Guard-3-8B (AI, 2024) as an automated safety classifier. For each prompt-response pair $(p_i, r_i)$, Llama-Guard evaluates the conversation against multiple safety categories including:

1. Illegal activities
2. Explicit content
3. Hate speech
4. Violence
5. Personal information disclosure
6. Harassment
7. Malicious code
8. Scams or fraud

A response is classified as "UNSAFE" if Llama-Guard detects violations in any category. The Attack Success Rate is computed as the proportion of responses classified as unsafe:

$$\text{ASR} = \frac{|\{(p_i, r_i) : \textsc{LlamaGuard}(p_i, r_i) = \text{UNSAFE}\}|}{N} \tag{8}$$

where $N = 520$ is the total number of AdvBench queries.

**Task Performance.** Evaluated using lm-evaluation-harness (Gao et al., 2024) with default settings for each benchmark.

# D Gradient Direction Analysis: Methodology

This appendix provides methodological details for the gradient direction analysis presented in Section 3.2.

### D.1 The High-Dimensionality Challenge

Computing cosine similarity between gradients and the reversion direction $\mathbf{r} = \theta_{\text{pretrain}} - \theta_{\text{aligned}}$ in billion-dimensional parameter spaces faces concentration of measure: vectors become nearly orthogonal regardless of true underlying alignment. Our preliminary experiments confirmed this issue, with full-parameter cosine similarities of approximately 0.006 even when significant correlations existed.

### D.2 TopK-Cosine Metric

To address this challenge, we introduce TopK-Cosine, which focuses on the parameters most modified during alignment training. For parameter subset $S$, let $I_k$ index the $k$ dimensions with largest $|r_j|$:

$$\text{TopK-Cos}(\mathbf{g}, \mathbf{r}; S, k) = \frac{\mathbf{g}_{I_k} \cdot \mathbf{r}_{I_k}}{\|\mathbf{g}_{I_k}\| \|\mathbf{r}_{I_k}\|} \quad (9)$$

This metric remains normalized (independent of gradient magnitude) while focusing on alignment-critical parameters. We use $k = 1000$ throughout our analysis.

Table 10 validates TopK-Cosine against standard cosine similarity for Qwen2.5-1.5B last-layer V projection. Both metrics yield similar correlation values ($r \approx 0.40$–$0.50$), confirming they capture the same underlying effect. However, TopK-Cosine provides approximately $4\times$ higher absolute values (0.119 vs 0.029), improving interpretability.

| Metric | HIGH | MOD | r | p |
|---|---|---|---|---|
| Full Cosine | 0.029 | 0.027 | 0.50 | $< 10^{-3}$ |
| TopK-Cosine | 0.119 | 0.104 | 0.41 | $< 10^{-3}$ |

Table 10: Comparison of TopK-Cosine and full cosine for Qwen2.5-1.5B last-layer V projection. Similar correlations confirm consistency; higher TopK-Cosine values improve interpretability.

### D.3 Experimental Protocol

**Reversion Direction.** For each aligned model, we compute $\mathbf{r} = \theta_{\text{pretrain}} - \theta_{\text{aligned}}$ using matched base models: Qwen2.5-1.5B for Qwen2.5-1.5B-Instruct and Llama-3.1-8B for Llama-3.1-8B-Instruct.

**Gradient Collection.** We compute per-sample gradients $\mathbf{g}_i = \nabla_\theta \mathcal{L}(x_i, y_i; \theta_{\text{aligned}})$ on 500 randomly sampled examples from Dolly, recording gradient norms $G_i = \|\mathbf{g}_i\|_2$.

**Parameter Subsets.** We partition parameters by layer position (last layer, middle third) and component type (Q/K/V/O projections, MLP). This enables localization of directional effects.

**Statistical Analysis.** Samples are stratified into HIGH (top 20%) and MODERATE (middle 20%) by gradient norm. We compute group means and Pearson correlation between gradient norm and TopK-Cosine for each parameter subset.

### D.4 Additional Parameter Subsets

Table 11 presents TopK-Cosine for additional parameter configurations not shown in the main paper.

For Qwen2.5-1.5B, all final-layer configurations show positive correlation, with V/O projections exhibiting the strongest signal. For LLaMA-3.1-8B, the MLP layer is the primary contributor to directional alignment, while attention projections show weak or no effect.

### D.5 Interpretation

The localization of directional effects to different components across architectures likely reflects differences in alignment training procedures. Qwen2.5's alignment appears to emphasize modifications to attention output transformations (V/O projections), while LLaMA's alignment concentrates in MLP layers. Despite this variation, the consistent pattern across both models is that directional alignment appears in final-layer parameters and is absent in middle layers, supporting the hypothesis that high-gradient samples reverse alignment-critical modifications.

## E Additional Experimental Results

### E.1 What Gets Filtered: Sample Audit

To characterize filtered samples and assess potential fairness concerns, we audit 10,000 Dolly samples using Qwen2.5-7B-Instruct, partitioning by gradient norm into three groups.

High-$G_i$ samples are format mismatches, not content-based outliers. They are dominated by short-answer tasks (classification 28.6%, closed QA 15.2%), averaging only 11.5 response tokens with high per-token loss (5.23). Large gradients arise because the aligned model's verbose output distribution diverges from terse targets—not because the content is semantically dissimilar. This indicates no fairness concern: filtering targets output format mismatch, not content or demographics.

| Model | Config | HIGH | MOD | r | p |
|---|---|---|---|---|---|
| Qwen2.5-1.5B | last1_MLP | 0.014 | 0.014 | 0.27 | $< 10^{-3}$ |
| | last1_QKVO | 0.212 | 0.191 | 0.36 | $< 10^{-3}$ |
| | last1_all | 0.044 | 0.043 | 0.34 | $< 10^{-3}$ |
| LLaMA-3.1-8B | last1_O | −0.098 | −0.092 | 0.09 | 0.22 |
| | last1_QKVO | −0.093 | −0.086 | −0.07 | 0.31 |
| | last1_all | −0.008 | −0.005 | −0.05 | 0.50 |

Table 11: TopK-Cosine for additional parameter configurations.

| Property | Low-$G_i$ | Moderate-$G_i$ | High-$G_i$ |
|---|---|---|---|
| Mean gradient norm | 1.35 | 3.89 | 16.77 |
| Mean loss | 1.67 | 2.33 | 5.23 |
| Mean answer tokens | 201.7 | 54.6 | 11.5 |

Table 12: Sample characteristics by gradient group. High-$G_i$ samples have short answers (11.5 tokens) and high per-token loss, indicating format mismatch rather than semantic dissimilarity.

### E.2 Task-Specific Performance by Checkpoint

**Overall trends.** Moderate-$G_i$ consistently mitigates catastrophic forgetting on GSM8K across continual learning stages. On Qwen3-4B, our method retains 67.4% after MedMCQA and 61.8% after SQuAD v2, compared to baselines' 44-53% and 38-40% respectively. MedMCQA training causes the most severe interference with GSM8K across all models, while Squad training shows variable effects, which sometimes recover performance (Qwen2.5-7B Random: 53.78 → 65.9%) and sometimes continue degradation (Qwen3-4B: 53 → 38%).

**Training stability.** Moderate-$G_i$ substantially reduces variance compared to Random sampling. The most striking example occurs on Qwen2.5-7B after MedMCQA: Random exhibits 11.1% standard deviation on GSM8K due to one seed collapsing to 40.9%, while Moderate-$G_i$ maintains 1.3% std with no catastrophic failures. This stability advantage persists across models and checkpoints, demonstrating that gradient-based selection produces more reliable continual learning dynamics.

### E.3 Extended Continual Learning Analysis

This section provides supplementary visualizations for the continual learning analysis in Section 5.4.

**Backward Transfer and Forgetting Measure Visualization.** Figure 5 visualizes the BWT and FM metrics from Table 8. Panel (a) shows Backward Transfer on Qwen3-4B, where Moderate-$G_i$'s advantage ($-4.3\%$ vs $-18.5\%$ baseline) is visually striking. Panel (b) reveals an important pattern: while methods show similar forgetting on Qwen2.5-7B (FM $\leq 3\%$), they diverge dramatically on Qwen3-4B (4.3% vs 18-20%). This suggests gradient-based selection becomes increasingly valuable as models become more susceptible to forgetting.

**Total Task Interference.** Figure 6 presents total task interference, the sum of all pairwise interference effects from Figure 3. On Qwen3-4B, Baseline accumulates $-37.1\%$ total interference compared to $-8.5\%$ for Moderate-$G_i$ ($4.4\times$ reduction). This cumulative view explains the performance gaps in Table 8: each training stage introduces less destructive interference with gradient-based selection, allowing knowledge to accumulate rather than cancel.

**Connection to Safety Alignment.** The metrics reinforce our central thesis: high-gradient samples cause both alignment drift and catastrophic forgetting through the same mechanism:

1. Large parameter updates reverse task-specific representations (causing forgetting)
2. Updates push parameters toward pretrained distributions (elastic reversion)
3. Extreme updates risk crossing safety basin boundaries (alignment collapse)

Moderate-$G_i$ addresses all three by preventing the high-gradient samples that trigger them. Consistent improvements across all metrics, especially where safety gains are largest, confirm that gradient-based selection taps into fundamental continual learning dynamics.

### E.4 Why Not Gradient Clipping?

A natural alternative to filtering high-gradient samples is gradient norm clipping, which bounds up-

| Model | Method | After GSM8K | | | After MedMCQA | | | After SQuAD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GSM8K | Med | SQuAD | GSM8K | Med | SQuAD | GSM8K | Med | SQuAD |
| Qwen2.5-7B | Random | 64.4 ± 1.6 | 57.5 ± 0.3 | 51.1 ± 0.8 | 53.8 ± 11.1 | 60.3 ± 0.5 | 56.4 ± 1.2 | 65.9 ± 3.0 | 59.6 ± 0.4 | 61.0 ± 4.7 |
| | Baseline | 67.1 ± 0.4 | 57.1 ± 0.3 | 50.3 ± 0.1 | 55.3 ± 1.3 | 61.1 ± 0.0 | 56.8 ± 1.1 | 64.5 ± 2.3 | 60.2 ± 0.1 | 59.3 ± 2.4 |
| | KL | 67.1 ± 0.4 | 56.8 ± 0.1 | 50.5 ± 0.2 | 70.8 ± 0.6 | 61.0 ± 0.2 | 57.4 ± 0.2 | 65.8 ± 4.6 | 60.5 ± 0.2 | 59.0 ± 3.3 |
| | O-LoRA | 56.6 ± 2.1 | 56.3 ± 0.0 | 51.1 ± 0.3 | 61.8 ± 0.8 | 61.0 ± 0.3 | 54.9 ± 0.9 | 40.4 ± 1.8 | 59.6 ± 0.2 | 59.1 ± 1.1 |
| | **Moderate-$G_i$** | 68.6 ± 1.8 | 56.7 ± 0.5 | 51.7 ± 0.7 | 65.9 ± 1.6 | 58.7 ± 0.2 | 57.3 ± 1.4 | 65.4 ± 4.0 | 58.2 ± 0.5 | 65.4 ± 5.6 |
| LLaMA-3.1-8B | Random | 19.8 ± 4.3 | 58.7 ± 0.4 | 54.6 ± 0.9 | 15.2 ± 0.9 | 60.0 ± 0.3 | 57.5 ± 2.1 | 22.4 ± 6.8 | 59.1 ± 0.4 | 74.2 ± 1.8 |
| | Baseline | 23.4 ± 1.2 | 59.0 ± 0.2 | 50.7 ± 0.3 | 15.0 ± 2.0 | 60.1 ± 0.4 | 56.2 ± 0.9 | 24.6 ± 1.7 | 59.4 ± 0.4 | 70.4 ± 2.4 |
| | KL | 23.8 ± 1.8 | 58.8 ± 0.2 | 51.0 ± 0.2 | 15.1 ± 1.5 | 59.7 ± 0.3 | 56.8 ± 0.5 | 22.1 ± 4.8 | 59.2 ± 0.2 | 70.7 ± 2.1 |
| | O-LoRA | 18.8 ± 2.2 | 59.3 ± 0.2 | 52.6 ± 0.5 | 22.7 ± 1.3 | 59.7 ± 0.1 | 54.6 ± 0.7 | 32.0 ± 2.1 | 59.4 ± 0.2 | 72.0 ± 0.3 |
| | **Moderate-$G_i$** | 24.0 ± 4.4 | 59.1 ± 0.2 | 52.5 ± 1.3 | 18.6 ± 3.8 | 58.3 ± 2.1 | 56.3 ± 1.5 | 20.7 ± 3.1 | 58.1 ± 1.8 | 70.2 ± 3.8 |
| Qwen3-4B | Random | 76.6 ± 2.6 | 55.2 ± 0.5 | 50.1 ± 0.0 | 53.4 ± 8.3 | 58.3 ± 0.2 | 50.4 ± 0.5 | 37.8 ± 7.3 | 57.5 ± 0.2 | 55.4 ± 4.6 |
| | Baseline | 76.6 ± 0.6 | 55.5 ± 0.2 | 50.1 ± 0.0 | 44.1 ± 2.3 | 59.4 ± 0.4 | 50.1 ± 0.0 | 40.4 ± 5.0 | 58.5 ± 0.0 | 52.9 ± 1.9 |
| | KL | 73.7 ± 2.7 | 55.5 ± 0.0 | 50.1 ± 0.0 | 47.7 ± 5.0 | 59.3 ± 0.4 | 50.2 ± 0.2 | 43.5 ± 4.3 | 58.5 ± 0.7 | 54.6 ± 6.9 |
| | O-LoRA | 77.2 ± 0.8 | 55.2 ± 0.1 | 50.1 ± 0.0 | 68.2 ± 3.2 | 59.1 ± 0.3 | 50.1 ± 0.0 | 52.9 ± 1.6 | 58.6 ± 0.1 | 58.2 ± 0.5 |
| | **Moderate-$G_i$** | 69.9 ± 1.0 | 54.7 ± 0.9 | 50.1 ± 0.0 | 67.4 ± 9.8 | 55.1 ± 0.3 | 50.2 ± 0.3 | 61.8 ± 8.7 | 54.7 ± 0.3 | 59.0 ± 5.7 |

Table 13: Task-specific performance across continual learning checkpoints for all methods. Each checkpoint evaluates on all three tasks (GSM8K, MedMCQA, Squad v2). Values show mean ± standard deviation across 3 seeds.

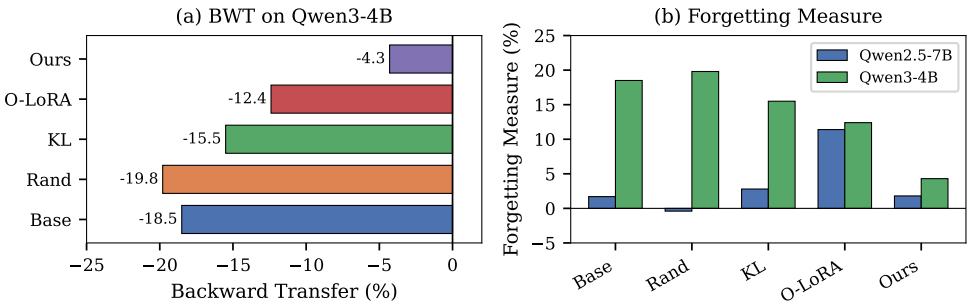


Figure 5: Continual learning metrics visualization. (a) BWT on Qwen3-4B shows 14.2% improvement for Moderate-$G_i$. (b) FM comparison reveals gradient-based selection provides the largest benefits on forgetting-prone models.

| Method | ASR↓ | TruthfulQA | Task Avg |
|---|---|---|---|
| Baseline (clip=1.0) | 36.7 | 38.2 | 59.1 |
| Clip 0.5 | 31.2 | 38.2 | 59.4 |
| Clip 0.1 | 32.4 | 37.3 | 58.9 |
| Moderate-$G_i$ | **10.2** | **42.5** | **60.9** |

Table 14: Gradient clipping vs. sample selection on Dolly with Qwen2.5-7B. Clipping provides only marginal ASR improvement (best: 31.2% at clip=0.5), while our method achieves 3× further reduction.

date magnitude without discarding data. We compare the two approaches on Dolly with Qwen2.5-7B (Table 14).

Gradient clipping provides only marginal improvement (best: 31.2% ASR at clip=0.5), while our method achieves 10.2%—a 3× further reduction. Clipping also fails to improve TruthfulQA (∼38% vs. our 42.5%). This reveals a key distinction: the problem is not gradient magnitude per se, but *which samples* generate those gradients. Clipping attenuates step size but still trains on high-gradient samples—the model still receives a learning signal pushing toward pretrained distributions. Our method removes these samples entirely, preventing their alignment-reversing content from influencing training.

### E.5 Task Order Robustness

To evaluate sensitivity to task ordering, we test two additional orderings and a domain substitution on LLaMA-3.1-8B (Table 15).

Under alternative orderings, our method achieves both the best safety (ASR 1.0% and 0.8%) and the best task performance (48.7% and 52.3%) on LLaMA-3.1-8B, demonstrating Pareto dominance. The cross-architecture variation observed in the original ordering likely reflects differences in how model families were aligned during post-training, rather than a fundamental limitation of

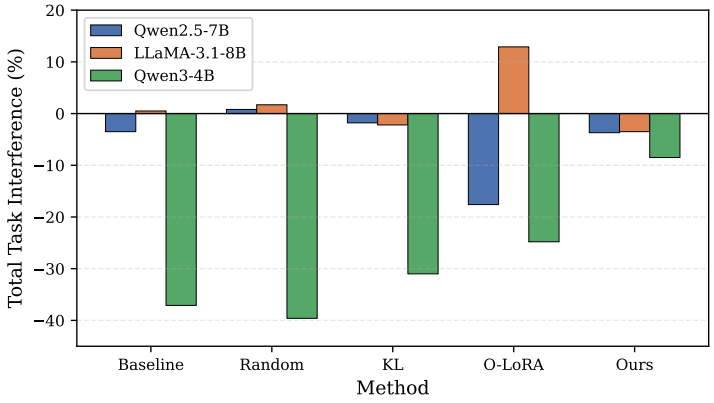


Figure 6: Total task interference (sum of all pairwise effects). Moderate-$G_i$ achieves $4.4\times$ reduction on Qwen3-4B.

| Method | ASR↓ | Task Avg |
|---|---|---|
| *Order 1: Dolly → MedMCQA → GSM8K → Squad* | | |
| Random | 7.8 | 45.9 |
| EWC | 15.5 | 47.3 |
| KL | 11.0 | 48.2 |
| O-LoRA | 1.2 | 48.4 |
| Moderate-$G_i$ | **1.0** | **48.7** |
| *Order 2: Dolly → Squad → MedMCQA → GSM8K* | | |
| Random | 15.5 | 51.6 |
| EWC | 12.8 | 50.7 |
| O-LoRA | 4.0 | 49.2 |
| Moderate-$G_i$ | **0.8** | **52.3** |
| *Domain substitution (Alpaca replacing Dolly), Qwen2.5-7B* | | |
| Baseline | 39.8 | 57.3 |
| Random | 19.0 | 59.1 |
| Moderate-$G_i$ | **10.4** | **59.6** |

Table 15: Task order robustness. Our method achieves best or near-best ASR and task accuracy across all configurations, demonstrating Pareto dominance independent of task ordering.

our approach. Results also generalize across initial datasets (Dolly → Alpaca).

### E.6 Computational Overhead

| Method | Time/Epoch | Relative Cost |
|---|---|---|
| Baseline | 14.4 min | 1.00× |
| KL Regularization | 20.4 min | 1.42× |
| O-LoRA | 14.8 min | 1.03× |
| Moderate-$G_i$ (ours) | 21.6 min | 1.5× |

Table 16: Wall-clock training time on Qwen2.5-7B on Dolly with single H100 GPU.

Our method adds 51% computational overhead due to gradient computation for candidate filtering. However, this overhead occurs only during training; inference cost is unchanged. The overhead can be reduced through gradient checkpointing, computing gradients only for LoRA parameters, and caching gradient statistics across epochs.

## F Limitations and Future Work

**Computational Overhead.** Our method requires additional gradient computation during training ($1.5\times$ baseline cost). While acceptable for most deployment scenarios, this overhead may be prohibitive for very large models or extremely limited compute budgets.

**Gradient Approximations.** We compute exact per-sample gradients, which requires sequential backward passes. Gradient approximation techniques such as influence functions or gradient sketching could reduce this cost but may affect selection quality.

**Dynamic Selection Threshold.** We use a fixed selection ratio $\rho = 0.2$ throughout training. Adaptive strategies that adjust selection strictness based on observed alignment drift could improve efficiency.

**Theoretical Characterization.** While we provide empirical evidence linking gradient magnitude to alignment drift, a formal theoretical analysis connecting sample gradients to safety basin geometry remains future work.

**Multi-Modal Models.** Our experiments focus on text-only models. Extending gradient-based selection to vision-language models or other modalities requires further investigation.