Safety LLMs fine-tuning via Conditional Information Bottleneck

Anonymous ACL submission

Abstract

Fine-tuning large language models on downstream tasks often degrades their safety alignment, a problem that compounds during sequential adaptation. We introduce Conditional Information Bottleneck (CIB), which preserves safety by encouraging fine-tuned representations to remain close to those of an aligned reference model. Our insight is that aligned models already encode safety-relevant structure, serving as implicit supervision without requiring safety labels. Information-theoretic analysis shows that maximizing mutual information between fine-tuned and reference representations preserves this structure, while the *alignment tax*—the performance cost of safety constraints—remains small for benign tasks where task labels are largely independent of safety structure. Experiments across multiple model families demonstrate substantial safety improvements with minimal performance degradation, and strong correlations between our theoretical quantities and harm rates validate our analysis.

1 Introduction

Large language models (LLMs) have achieved remarkable capabilities across diverse tasks, from reasoning and coding to creative writing and scientific analysis. This success has shifted the machine learning paradigm from training task-specific models toward efficiently adapting powerful pre-trained foundations. However, full fine-tuning of billion-parameter models remains computationally prohibitive for most practitioners, motivating the development of parameter-efficient alternatives. Parameter-efficient fine-tuning (PEFT) methods have emerged as practical solutions for adapting LLMs to downstream tasks with minimal parameter updates, including adapters (Houlsby et al., 2019), prefix/prompt tuning (Li and Liang, 2021; Lester et al., 2021), and Low-Rank Adaptation (LoRA) (Hu et al., 2021). These methods impose *explicit structural constraints*: adapters insert small bottleneck modules, prompt tuning restricts learning to continuous embeddings, and LoRA confines updates to low-rank subspaces. Such constraints create deliberate information bottlenecks that align naturally with the Information Bottleneck (IB) principle (Tishby et al., 2000)

$$\min_{p(Z|X)} I(X;Z) - \beta I(Z;Y), \tag{1}$$

which seeks representations Z that compress input X while preserving information about target Y.

Recent work has successfully applied IB principles to various LLM challenges, including reducing memorization (Wang et al., 2024), improving reasoning (Lei et al., 2024), and defending against jailbreak attacks (Liu et al., 2024). Yet these primarily address specific applications or input-level perturbations. A critical question remains:

How can we leverage IB principles to preserve safety alignment during fine-tuning itself?

This question is urgent because even fine-tuning on benign tasks (e.g., mathematics, coding) can degrade safety alignment (Qi et al., 2023; Yang et al., 2023), making models vulnerable to harmful requests. Standard fine-tuning provides no mechanism to distinguish safety-relevant features from task-specific features during compression. This problem intensifies during sequential adaptation across multiple tasks, where safety erodes progressively. Existing solutions either require explicit safety labels (Hsu et al., 2024) or apply uniform regularization that fails to selectively preserve safety-relevant information (Rosati et al., 2024).

We address this via the Conditional Information Bottleneck (CIB) principle (Fischer, 2020). Our key insight: aligned pre-trained models already encode safety in their representations, providing implicit supervision without requiring safety labels. We condition fine-tuning on the reference model

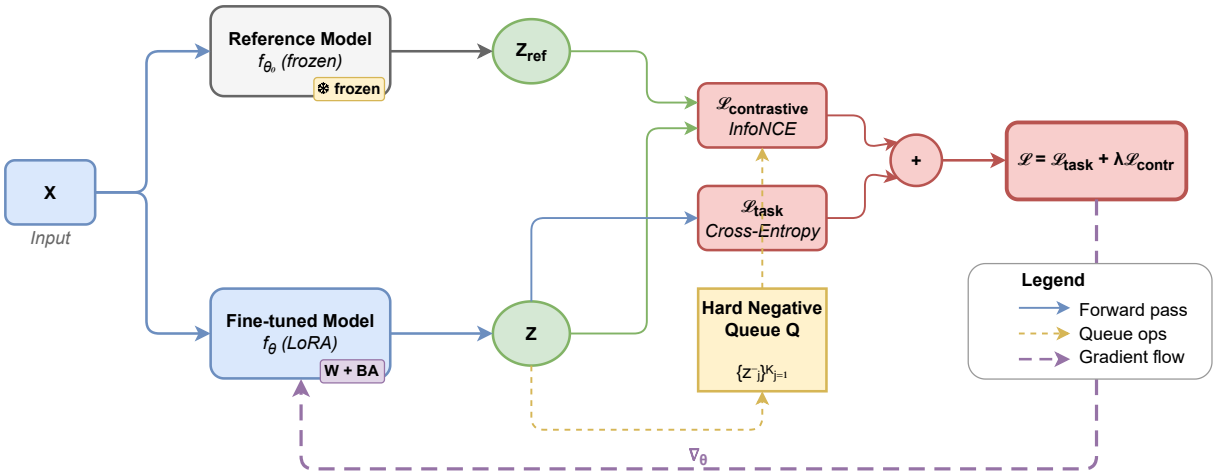

Figure 1: **CIB Framework Overview.** Input X passes through both the frozen reference model f_{θ_0} and the LoRA-adapted model f_θ. The contrastive loss maximizes mutual information $I(Z; Z_{\text{ref}})$ between representations, preserving safety-relevant structure from the aligned reference. Hard negatives are mined based on representation divergence and task loss ($h_i = (1 - \cos(z_i, z_{\text{ref},i})) \cdot \ell_i$). The combined objective $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\mathcal{L}_{\text{contrastive}}$ updates only LoRA parameters while maintaining safety alignment.

f_{θ_0}:

$$\min_{p(Z|X,C)} I(X;Z|C) - \beta I(Z;Y|C), \quad C = f_{\theta_0}(X). \tag{2}$$

Besides, a natural concern is whether preserving safety compromises task performance. We formalize this trade-off as the *alignment tax*: the reduction in task-relevant information due to safety constraints. Our analysis provides two results: (1) for deterministic representations, conditioning reduces compression by exactly $I(Z;C)$ (Theorem 1), directly motivating maximizing mutual information with the reference model; (2) the alignment tax is bounded by $I(Y;C)$ (Proposition 1), the mutual information between task labels and reference representations. For benign tasks—where correct answers do not depend on circumventing safety mechanisms—$I(Y;C)$ is empirically small, explaining why CIB preserves safety with minimal performance cost.

While prior work applies IB to jailbreak defense (Liu et al., 2024) or general LLM tasks (Wang et al., 2024; Lei et al., 2024), our contribution is an information-theoretic framework specifically for safety-preserving fine-tuning, complementing geometric perspectives (Safe LoRA (Hsu et al., 2024)) and gradient-based analyses (SafeGrad (Wu et al., 2025)).

Key Contributions.

1. An information-theoretic analysis of safety-preserving fine-tuning, establishing compression decomposition and bounding the alignment tax.
2. A practical contrastive learning algorithm that conditions on reference model representations without requiring safety labels.
3. Empirical validation showing substantial safety improvements with minimal performance cost and strong correlation with theoretical predictions.

Organization. Section 2 establishes background and formalizes the problem. Section 3 develops the analysis for our method. Section 4 derives the algorithm. Section 5 presents empirical validation.

2 Preliminaries

We establish the information-theoretic background and formalize the problem setup for safety-preserving LLM fine-tuning.

2.1 Information Bottleneck Background

Let X, Y, and Z denote random variables representing inputs, targets, and learned representations, respectively. We denote the entropy of X as $H(X)$, the conditional entropy of X given Y as $H(X|Y)$, and the mutual information between X and Y as:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{3}$$

Mutual information quantifies the amount of information that one variable provides about another.

The Information Bottleneck (IB) principle (Tishby et al., 2000) provides an information-theoretic framework for learning compressed representations. Given input X and target Y, IB seeks a representation Z that solves:

$$\min_{p(Z|X)} I(X;Z) - \beta I(Z;Y), \tag{4}$$

where $\beta > 0$ controls the trade-off between compression (minimizing $I(X;Z)$) and prediction (maximizing $I(Z;Y)$). In deep learning, representations are typically deterministic functions: $Z = f_\theta(X)$. In this case, $H(Z|X) = 0$, and the compression term $I(X;Z)$ simplifies to $H(Z)$, encouraging compact representations.

2.2 Conditional Information Bottleneck

While IB provides a principled framework for compression, it lacks a mechanism to direct compression toward preserving specific properties. The Conditional Information Bottleneck (CIB) (Fischer, 2020) extends IB by introducing a conditioning variable C that guides the compression process:

$$\min_{p(Z|X,C)} I(X;Z|C) - \beta I(Z;Y|C). \tag{5}$$

The key distinction from standard IB is the conditioning on C in both terms:

- **Conditional Compression** $I(X;Z|C)$: Measures the information Z retains about X beyond what is explained by C. Minimizing this encourages Z to depend on X only through C-relevant features.
- **Conditional Prediction** $I(Z;Y|C)$: Measures the predictive information Z provides about Y given C. Maximizing this ensures Z remains useful for the task within each conditioning context.

2.3 Problem Formulation

We formalize fine-tuning a pre-trained LLM with parameters θ_0 on a downstream task while preserving safety alignment. Let $f_{\theta_0} : \mathcal{X} \to \mathcal{Z}$ denote the pre-trained model's representation mapping, and $f_\theta : \mathcal{X} \to \mathcal{Z}$ denote the fine-tuned model.

Definition 1 (CIB Fine-tuning)**.** *Given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, and reference model f_{θ_0} providing conditioning via representations $Z_{ref} = f_{\theta_0}(X)$, we learn parameters θ minimizing:*

$$\min_\theta I(X;Z|Z_{ref}) - \beta I(Z;Y|Z_{ref}), \quad Z = f_\theta(X). \tag{6}$$

We employ LoRA (Hu et al., 2021) for parameter-efficient adaptation:

$$W_l \leftarrow W_l + B_l A_l, \quad B_l \in \mathbb{R}^{d\times r}, A_l \in \mathbb{R}^{r\times k}, \quad r \ll \min(d,k), \tag{7}$$

where $\mathcal{L}$ denotes the adapted layers. This low-rank constraint provides implicit regularization, which we complement with explicit CIB-guided training.

3 Analysis

We establish two key results: (1) conditioning reduces compression by exactly $I(Z;C)$ (Section 3.1), motivating our contrastive learning approach; and (2) a bound on the *alignment tax*—the performance cost of safety constraints—showing it is bounded by $I(Y;C)$ and empirically small for benign tasks (Section 3.2).

3.1 Compression Reduction under Conditioning

For deterministic representations, conditioning on C reduces compression by exactly $I(Z;C)$. This decomposition directly motivates maximizing $I(Z;C)$ to minimize conditional compression $I(X;Z|C)$.

Theorem 1 (Compression Reduction)**.** *For deterministic representations $Z = f(X)$:*

$$I(X;Z|C) = I(X;Z) - I(Z;C). \tag{8}$$

Proof. See Appendix A.1 □

Theorem 1 follows from standard information-theoretic identities; its value lies not in mathematical novelty but in its direct prescription for algorithm design: to minimize conditional compression, maximize $I(Z;C)$ with the reference model. The compression gap $I(Z;C)$ measures how much C explains about Z. This quantity is larger when fine-tuned representations preserve structure from the conditioning variable, directly motivating maximizing $I(Z;C)$ to minimize $I(X;Z|C)$. Note that $0 \le I(Z;C) \le \min\{H(Z), H(C)\}$.

3.2 Alignment Tax: Performance Cost of Safety

When fine-tuning on a downstream task while preserving safety, does enforcing safety constraints hurt task performance? We formalize this trade-off as the alignment tax.

Definition 2 (Alignment Tax)**.** *Consider fine-tuning on task labels Y while conditioning on*

safety-relevant structure C. The alignment tax is the reduction in task-relevant information:

$$\textit{Tax}(Y, C) = I(Z; Y) - I(Z; Y|C). \quad (9)$$

The alignment tax measures how much representation Z loses in predictive power for Y when conditioning on safety constraints C. A large tax indicates conflict; a small tax suggests compatibility.

Proposition 1 (Alignment Tax Bound)**.** *The alignment tax is bounded by:*

$$\textit{Tax}(Y, C) \leq I(Y; C). \quad (10)$$

Proof. See Appendix A.2 □

When is the alignment tax small? The bound shows the tax is small when $I(Y; C)$ is small, indicating that task correctness and safety are highly independent. For benign downstream tasks (mathematics, coding, medical QA), the correct answer does not depend on safety constraints. A mathematical solution is correct or incorrect independent of safety properties, implying small $I(Y; C)$. Recent work supports this: (Wu et al., 2025) observes non-conflicting task and safety gradients on benign data, while (Kim et al., 2025) shows safety degradation stems from optimization issues rather than fundamental task-safety conflicts. Additionally, well-aligned pre-training ensures safety constraints are implicitly satisfied by competent solutions, avoiding new conflicts during fine-tuning. Section 4 operationalizes these insights through contrastive learning that maximizes $I(Z; C)$.

Identifying High-Tax Samples. While the population-level $I(Y; C)$ is small for benign tasks, individual samples may still incur an alignment tax. Samples where the fine-tuned representation Z diverges from the reference C while exhibiting high task loss are candidates in which task adaptation conflicts with the safety structure. This motivates our hard negative mining strategy (Section 4.4): by identifying samples with both high divergence and high task loss, we focus contrastive learning on the decision boundary where safety-task tensions are most pronounced.

4 Method: CIB-Guided Fine-tuning

We translate our theoretical insights into a practical algorithm. Section 3 established that maximizing $I(Z; C)$ minimizes conditional compression $I(X; Z|C)$ and reduces the alignment tax. We operationalize this through contrastive learning.

4.1 Algorithm

From Theorem 1, minimizing $I(X; Z|C)$ for deterministic representations is equivalent to:

$$\min I(X; Z|C) = \min I(X; Z) - \max I(Z; C). \quad (11)$$

Since LoRA constrains representation capacity, the key optimization target becomes maximizing $I(Z; C)$.

We maximize $I(Z; C)$ via contrastive learning. Following (Oord et al., 2018), the InfoNCE objective provides a tractable lower bound:

$$I(Z; C) \geq \log(K + 1) - \mathcal{L}_{\text{InfoNCE}}, \quad (12)$$

where

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{(z, z^+)} \log \frac{e^{s(z, z^+)/\tau}}{e^{s(z, z^+)/\tau} + \sum_{j=1}^{K} e^{s(z, z_j^-)/\tau}} \quad (13)$$

z_i^+ shares the same C value as z_i, and $\{z_j^-\}_{j=1}^K$ have different C values, and $s(z, z') = \cos(z, z')$ is cosine similarity.

4.2 Overall Training Objective

The CIB-guided fine-tuning objective combines task performance with safety preservation:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{contrastive}}, \quad (14)$$

where $\mathcal{L}_{\text{task}} = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i|x_i)$ optimizes task performance, $\mathcal{L}_{\text{contrastive}}$ maximizes $I(Z; C)$, and $\lambda > 0$ controls regularization strength.

4.3 Safety-Preserving Contrastive Loss

Reference representations as conditioning. Let $Z_{\text{ref}} = f_{\theta_0}(X)$ denote representations from the aligned reference model's last hidden layer, and $Z = f_\theta(X)$ denote fine-tuned representations. Our contrastive objective maximizes $I(Z; Z_{\text{ref}})$:

$$\mathcal{L}_{\text{safety}} = -\mathbb{E}_{(z, z_{\text{ref}})} \log \frac{e^{s(z, z_{\text{ref}})/\tau}}{e^{s(z, z_{\text{ref}})/\tau} + \sum_{j=1}^{K} e^{s(z, z_j^-)/\tau}}. \quad (15)$$

where $(z_i, z_{\text{ref},i})$ forms a positive pair (same input, different models), $\{z_j^-\}_{j=1}^K$ are negatives sampled from a queue, and $s(z, z') = \cos(z, z')$ is cosine similarity.

We implement conditioning not through explicit conditional distributions $p(Z|X, C)$, but through a regularization that encourages Z to preserve structure from C. This soft conditioning via contrastive learning is analogous to how knowledge distillation implements implicit conditioning on teacher representations.

4.4 Hard Negative Mining.

Section 3.2 showed that alignment tax concentrates on samples where task adaptation conflicts with safety-relevant structure. We operationalize this insight by defining sample hardness:

$$h_i = (1 - \cos(z_i, z_{\text{ref},i})) \cdot \ell_i, \tag{16}$$

where the first term measures representation divergence from the reference (potential safety drift) and $\ell_i = -\log p_\theta(y_i|x_i)$ measures task difficulty. Samples with high h_i exhibit both properties: the model struggles on the task *and* its representations deviate from safety-aligned structure. Using these as hard negatives focuses contrastive learning on the decision boundary where safety-task conflicts are most pronounced. We maintain a priority queue $\mathcal{Q}$ of size M and sample top-K candidates, following semi-hard mining principles (Schroff et al., 2015) to avoid destabilizing gradients from extremely hard negatives.

5 Experiments

We evaluate CIB on safety-preserving fine-tuning across single-task and sequential learning scenarios. Our experiments investigate: (1) whether CIB preserves safety during task-specific fine-tuning, (2) whether it maintains alignment throughout continual domain adaptation, and (3) whether empirical improvements correlate with $I(Z;C)$ as predicted by theory.

Models. We evaluate on three model families: Llama-2-7B-Chat, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Qwen3-4B-Instruct (Yang et al., 2025), all with RLHF-based safety alignment.

CIB Hyperparameters. We use contrastive weight $\lambda = 5.0$, temperature $\tau = 0.2$, queue size $M = 512$, and $K = 64$ negatives per batch.

5.1 Safety Preservation During Fine-Tuning

Setup. Following (Qi et al., 2023), we fine-tune aligned models on Dolly (Conover et al., 2023), a 15,000-sample instruction dataset, using LoRA with rank $r = 32$, $\alpha = 64$ for one epoch.

Baseline. We compare against standard LoRA and three safety-preserving fine-tuning baselines: Vaccine (Huang et al., 2024d) (adversarial perturbations), Safe LoRA (Hsu et al., 2024) (subspace projection), and SaLoRA (Li et al., 2025b) (orthogonal safety module). The "Base" row shows pre-aligned models without task fine-tuning as a reference.

Metrics. We measure (1) **Eval**: cross-entropy loss on Dolly test set (task adaptation quality); (2) **HRR**[1]: harmful response rate on AdvBench adversarial prompts (Zou et al., 2023) (safety); (3) zero-shot capabilities on ARC-Challenge, GSM8K, ToxiGen, and TruthfulQA.

Results. CIB achieves the strongest safety across model families. On Llama-3.1-8B, CIB reaches 4.8% HRR versus 8.1% for the next-best method (SaLoRA), while maintaining task performance (Eval: 1.3 vs 1.2) and leading on ARC-Challenge (52.9). On Qwen-2.5-7B, CIB reduces HRR to 1.5% (versus 3.4% for SaLoRA) while achieving best performance on ARC-Challenge (55.3) and GSM8K (79.9). For Llama-2-7B, CIB matches perfect safety (0.0% HRR) with comparable capabilities. Standard LoRA consistently shows best task adaptation (Eval) but severe safety degradation (16.7–25.5% HRR), confirming the need for safety-preserving methods.

Alignment Tax Validation. Proposition 1 predicts alignment tax is bounded by $I(Y;C)$, which is small when task correctness and safety are independent. Table 1 confirms this: CIB reduces HRR from 25.5% to 4.8% on Llama-3.1-8B, a 20.7 percentage point reduction. Across all models, CIB maintains $> 95\%$ of standard LoRA's task performance while improving safety by 3-8×, validating that $I(Y;C)$ is negligible for benign tasks.

5.2 Continual Safety Alignment

Setup. We evaluate sequential adaptation where models fine-tune continuously across tasks: Dolly (general instructions) → GSM8K (mathematics) → MedMCQA (medical QA) → SQuAD v2 (reading comprehension). We begin with Dolly to reduce initial over-refusal behavior common in RLHF-aligned models, enabling more reliable safety evaluation on subsequent tasks. We acknowledge this ordering may favor methods that preserve the post-Dolly safety level. LoRA weights are updated sequentially without resetting between tasks (rank 32, $\alpha = 64$, lr=1×10^{-4}, 1 epoch per task). We

[1]Harmful Response Rate (HRR) is evaluated using Llama-Guard-3-8B (Inan et al., 2023) as an automated safety classifier. See Appendix C.6 for details.

Methods	Eval ↓	HRR ↓	ARC-C	GSM8K	ToxiGen	TruthfulQA
Llama-3.1-8B						
Base	1.9	1.4	52.0	75.2	53.3	45.5
LoRA	**1.2**	25.5	51.2	72.4	44.9	39.0
Vaccine	1.3	21.3	44.3	39.5	43.4	34.1
Safe LoRA	1.3	11.0	51.1	75.6	48.7	42.0
SaLoRA	**1.2**	8.1	52.3	**75.7**	**49.3**	41.8
CIB (Ours)	1.3	**4.8**	**52.9**	74.8	48.4	**42.7**
Qwen-2.5-7B						
Base	3.6	0.0	53.0	76.4	57.2	56.3
LoRA	**1.2**	24.7	55.0	60.2	57.2	44.5
Vaccine	**1.2**	19.3	54.6	74.3	**57.9**	44.5
SaLoRA	**1.2**	3.4	55.0	69.5	57.2	49.2
CIB (Ours)	1.3	**1.5**	**55.3**	**79.9**	57.3	**50.9**
Llama-2-7B						
Base	2.5	0.0	43.3	20.1	52.9	37.2
LoRA	**1.1**	21.4	44.4	19.6	44.7	32.3
Vaccine	**1.1**	16.7	42.6	11.6	41.1	31.7
Safe LoRA	1.2	**0.0**	45.6	21.5	43.8	33.1
SaLoRA	**1.1**	**0.0**	**45.9**	**23.6**	**49.5**	**34.7**
CIB (Ours)	1.2	**0.0**	**45.9**	20.9	47.8	32.3

Table 1: Safety-preserving fine-tuning results. CIB achieves superior safety (lowest HRR among fine-tuned models) with competitive task performance. Base shows pre-aligned models without task fine-tuning, serving as a reference for the safety ceiling (lowest achievable HRR) and capability baseline. The meaningful comparison is among fine-tuned methods (LoRA through CIB). Bold indicates the best fine-tuned model performance. Single-task experiments use fixed random seeds for reproducibility. Table 2 reports the mean ± std over 3 runs for continual learning due to higher variance from sequential training.

compare against standard LoRA, random sample selection, KL regularization, and O-LoRA (Wang et al., 2023). The task order begins with Dolly to reduce initial refusal behavior, enabling more reliable safety evaluation on subsequent tasks.

Metrics. We measure safety (HRR on AdvBench), truthfulness (TruthfulQA), general capabilities (ARC-Challenge, BoolQ, HellaSwag, Winogrande), and backward transfer (BT: average accuracy on GSM8K, MedMCQA, and SQuAD v2 evaluated after completing all four tasks; Dolly excluded as a generation task).

Results. CIB achieves the lowest HRR across all models: 4.7% (Qwen-2.5-7B), 4.9% (Llama-3.1-8B), and 1.4% (Qwen-3-4B), representing $7.8\times$, $9.0\times$, and $11.9\times$ reductions versus standard LoRA, and $3.5\times$, $4.8\times$, and $4.5\times$ improvements over O-LoRA. This validates that maximizing $I(Z;C)$ throughout sequential adaptation preserves safety-relevant structure even as models acquire new task knowledge.

For continual learning, CIB maintains competitive backward transfer (51.4-58.6%) and achieves the highest ARC-Challenge scores across all models (57.5-60.7%). On TruthfulQA, CIB performs comparably to or better than baselines, with notable gains on Qwen-2.5-7B (47.7% vs 38.2-42.9%). Performance on BoolQ, HellaSwag, and Winogrande remains competitive, demonstrating that safety preservation does not compromise general capabilities.

5.3 Validating the Information-Theoretic Framework

Theorem 1 predicts that maximizing $I(Z;C)$ reduces conditional compression $I(X;Z|C)$, preserving safety-relevant information. We validate this by analyzing the relationship between $I(Z;C)$ and safety across contrastive weights $\lambda \in \{1, 2, 5, 10\}$.

Measuring $I(Z;C)$. We estimate $I(Z;C)$ using the InfoNCE lower bound (Oord et al., 2018):

$$I(Z;C) \geq \log(K+1) - \mathcal{L}_{\text{contrastive}}, \qquad (17)$$

where $K = 512$ negative samples. We compute this bound after training for each (model, λ) configuration.

Method	HRR↓	TruthQA	BT↑	ARC-C	BoolQ	HellaSwag	WinoG
			Qwen-2.5-7B				
LoRA	36.7 ± 13.6	38.2 ± 1.2	61.3 ± 0.9	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	**62.2** ± 2.6	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	61.7 ± 2.1	**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	53.0 ± 0.8	56.6 ± 0.9	86.1 ± 1.2	79.3 ± 0.4	71.8 ± 0.6
CIB	**4.7** ± 2.9	**47.7** ± 0.7	58.6 ± 3.4	57.5 ± 0.8	84.3 ± 1.6	**80.4** ± 0.2	69.5 ± 0.5
			Llama-3.1-8B				
LoRA	44.2 ± 22.5	37.8 ± 0.7	51.5 ± 1.0	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	51.9 ± 1.9	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	50.7 ± 2.1	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	**54.5** ± 0.9	56.4 ± 1.2	84.7 ± 0.6	78.3 ± 0.6	**74.0** ± 0.7
CIB	**4.9** ± 2.3	39.4 ± 0.7	51.4 ± 1.6	**58.1** ± 0.4	84.4 ± 0.4	**78.6** ± 0.0	73.9 ± 0.5
			Qwen-3-4B				
LoRA	16.6 ± 7.7	39.7 ± 0.7	50.6 ± 1.0	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	50.2 ± 3.9	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	52.2 ± 2.4	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	**56.6** ± 0.4	59.5 ± 1.3	85.2 ± 0.8	70.7 ± 1.1	68.7 ± 0.7
CIB	**1.4** ± 1.7	41.6 ± 0.5	55.4 ± 2.9	**60.7** ± 0.5	**86.6** ± 0.3	69.0 ± 0.1	66.9 ± 0.4

Table 2: Continual safety alignment results. Models fine-tune sequentially (Dolly → GSM8K → MedMCQA → SQuAD v2) with LoRA weights updated continuously. BT measures backward transfer (average accuracy on GSM8K, MedMCQA, SQuAD v2 after all tasks). CIB achieves superior safety with competitive, continual learning performance.

Model	λ	$I(Z;C)\uparrow$	HRR(%)↓	Eval↓	Utility
Llama-3.1-8B	1.0	2.98	15.4	1.2	62.2
	2.0	3.05	6.7	1.2	62.1
	5.0	3.10	4.8	1.3	62.4
	10.0	3.13	2.1	1.4	62.0
Llama-2-7B	1.0	2.87	0.8	1.1	59.1
	2.0	2.90	0.0	1.1	59.0
	5.0	2.94	0.0	1.2	59.1
	10.0	2.97	0.0	1.3	58.6
Qwen2.5-7B	1.0	2.87	5.0	1.2	62.8
	2.0	2.97	2.5	1.2	62.4
	5.0	3.03	1.5	1.3	62.6
	10.0	3.06	1.2	1.5	62.4

Table 3: Ablation on contrastive weight λ. Higher $I(Z;C)$ correlates strongly with lower HRR (Llama-3.1: $r = -0.98$; Qwen2.5: $r = -0.99$) while utility remains stable across λ values.

Results. Figure 2(a) shows $I(Z;C)$ increases monotonically with λ across all models, confirming our contrastive objective successfully maximizes mutual information with the reference model. Panel (b) shows corresponding HRR decreases. Critically, panel (c) reveals strong within-model negative correlations: Llama-3.1-8B ($r = -0.98$, $p < 0.05$) and Qwen2.5-7B ($r = -0.99$, $p < 0.05$). Llama-2-7B saturates at 0% HRR for $\lambda \geq 2$, indicating complete safety preservation with modest $I(Z;C)$ increases. These correlations are consistent with Theorem 1's prediction that $I(Z;C)$ and safety preservation are linked, though we note correlation does not establish that the contrastive objective is the unique or optimal way to achieve this effect.

Table 3 [2] shows utility remains stable across λ values (variation < 0.5 points), confirming Proposition 1: the alignment tax $I(Y;C)$ is negligible for benign tasks. Despite a narrow $I(Z;C)$ range (2.87-3.13 nats), small increases yield substantial safety gains, where Llama-3.1-8B achieves 13.3 point HRR reduction with only $\Delta I(Z;C) = 0.15$ nats, demonstrating the efficiency of our approach.

We conduct ablations to understand the contribution of each component.

Effect of Contrastive Weight λ We investigate the impact of contrastive weight λ on the safety-performance trade-off. Increasing λ substantially reduces harmful response rates while minimally affecting evaluation quality or general capabilities. For Llama-3.1-8B, $\lambda = 10$ achieves 7.3× lower HRR (2.1% vs 15.4%) compared to $\lambda = 1$, with only marginal Eval degradation (1.2→1.4). Llama-2-7B reaches 0% HRR at $\lambda \geq 2$, while Qwen2.5-7B reduces HRR from 5.0% to 1.2%. Critically, average utility remains stable across all settings

[2] Llama-2-7B saturates at 0% HRR for $\lambda \geq 2$, limiting correlation interpretation.

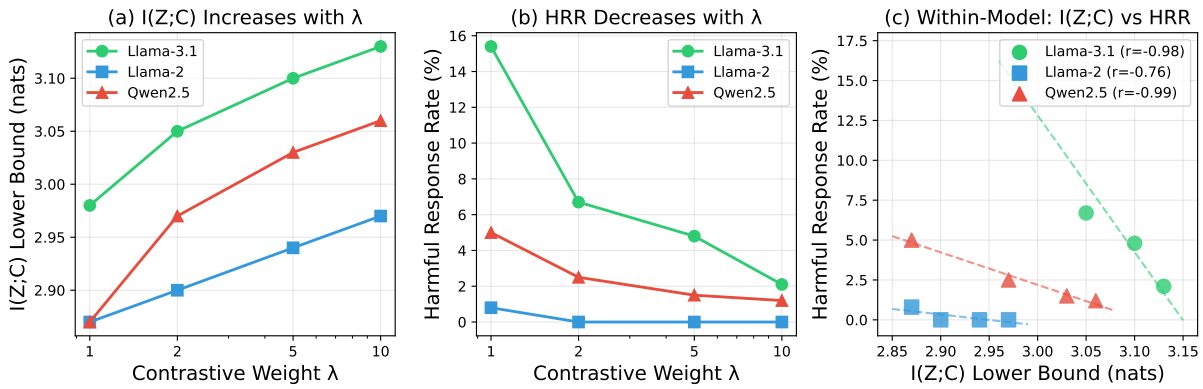

Figure 2: Empirical validation of the information-theoretic framework. (a) $I(Z;C)$ increases monotonically with contrastive weight λ. (b) Harmful response rate decreases with λ. (c) Within-model correlation between $I(Z;C)$ and HRR shows strong negative relationships (Llama-3.1: $r = -0.98$; Qwen2.5: $r = -0.99$), validating Theorem 1's prediction that maximizing $I(Z;C)$ preserves safety.

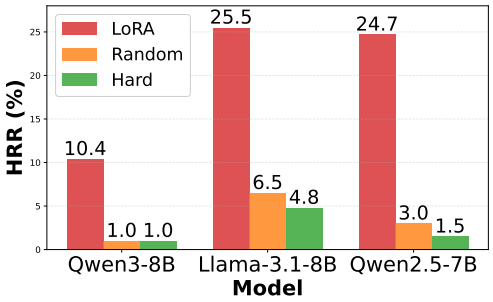

Figure 3: Comparison of negative sampling strategies. Hard negative mining selects negatives with the highest similarity to positives, achieving substantially lower harmful response rates compared to random negative selection. Both contrastive approaches dramatically outperform baseline LoRA across all models.

(within ± 0.4 points), demonstrating that stronger safety constraints preserve general capabilities. We use $\lambda = 5$ as the default, balancing safety improvement and model quality.

Hard Negative Mining We compare three negative sampling strategies: (1) baseline LoRA without contrastive learning, (2) random negative selection from the queue (Normal), and (3) hard negative mining that selects negatives with the highest similarity to positives (Hard). Figure 3 shows harmful response rates across three models. Hard negative mining substantially improves safety over random selection. On Llama-3.1-8B, hard negatives reduce HRR from 6.5% to 4.8% (26% relative improvement). Qwen2.5-7B shows the strongest effect, achieving 1.5% HRR with hard negatives versus 3.0% with random selection (50% reduction). Both contrastive approaches dramatically outperform baseline LoRA, which exhibits HRR of $10.4 - 25.5\%$ depending on the model. The results demonstrate that mining informative negatives, those close to the decision boundary, enables more effective safety learning than random sampling.

6 Conclusion

We introduced Conditional Information Bottleneck (CIB) for safety-preserving LLM fine-tuning. Our key insight is that aligned models encode safety in their representations, allowing us to preserve safety by conditioning on reference model outputs without requiring explicit safety labels. We established theoretical guarantees connecting mutual information maximization to safety preservation and bounded the alignment tax for benign tasks. Experiments across multiple model families validated our framework, showing substantial safety improvements with minimal performance cost in both single-task and continual learning settings.

Limitations. CIB requires reference model access during fine-tuning, limiting API-only scenarios. The hard negative queue introduces memory overhead. While effective against benign degradation, CIB does not defend against adversarial fine-tuning with explicitly harmful data.

References

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, Alexey Frolov, and Kirill Andreev. 2023. Information bottleneck analysis of deep neural networks via lossy compression. *arXiv preprint arXiv:2305.08013*.

Stephen Casper and 1 others. 2024. Improving large language model safety with contrastive representation learning. *arXiv preprint arXiv:2506.11938*.

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. *Databricks Blog*.

Ian Fischer. 2020. The conditional entropy bottleneck. *Entropy*, 22(9):999.

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*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

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, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024a. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*.

Kaifeng Huang, Yifan Zhai, Zhengyang Wu, Jinghan He, and Peter Henderson. 2024b. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. In *Conference on Neural Information Processing Systems*.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Tekin, and Ling Liu. 2024c. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:104521–104555.

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

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rber, Yasmine Bhosale, Somesh Teli, Robin Vestman, Jessica Caballero, Elena Kochkina, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Minseon Kim and 1 others. 2025. Rethinking safety in llm fine-tuning: An optimization perspective. *arXiv preprint arXiv:2508.12531*.

Shiye Lei, Chujie Huang, Junda Mu, and Dongkuan Lyu. 2024. Revisiting llm reasoning via information bottleneck. *arXiv preprint arXiv:2507.18391*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025a. Finding and reactivating post-trained llms' hidden safety mechanisms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025b. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yan-Shuo Liang, Jiarui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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*.

Zichuan Liu, Zefan Wang, Linjie Xu, Jinyu Wang, Lei Song, Tianchun Wang, Chunlin Chen, Wei Cheng, and Jiang Bian. 2024. Protecting your llms with information bottleneck. In *Advances in Neural Information Processing Systems*.

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*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Domenic Rosati, Leila Wehbe, Alex Tamkin, Dylan Hadfield-Menell, Noah Goodman, and Jacob Steinhardt. 2024. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.

Changsheng Wang, Zichuan Liu, Weizhe Chen, and Zibin Zeng. 2024. Breaking memorization barriers in llm code fine-tuning via information bottleneck for improved generalization. *arXiv preprint arXiv:2510.16022*.

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.

Zonghan Wu and 1 others. 2025. Safegrad: Gradient surgery for safe llm fine-tuning. *arXiv preprint arXiv:2508.07172*.

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*.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Improved few-shot jailbreaking can circumvent aligned language models and their defenses, 2024. *URL https://arxiv. org/abs/2406.01288*.

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, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. *arXiv preprint arXiv:2406.04313*.

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 Proofs

A.1 Proof of Theorem 1

By the chain rule for mutual information:

$$I(X, C; Z) = I(X; Z) + I(C; Z|X) = I(C; Z) + I(X; Z|C). \tag{18}$$

Rearranging yields $I(X; Z|C) = I(X; Z) + I(C; Z|X) - I(C; Z)$. For deterministic representations $Z = f(X)$, we have $H(Z|X) = 0$, which implies $I(C; Z|X) = H(Z|X) - H(Z|X, C) = 0$. Therefore:

$$I(X; Z|C) = I(X; Z) - I(Z; C). \tag{19}$$

A.2 Proof of Proposition 1

We decompose the alignment tax using the chain rule. By definition:

$$\begin{aligned}
\text{Tax}(Y, C) &= I(Z; Y) - I(Z; Y|C) && (20)\\
&= I(Z; Y, C) - I(Z; C) - I(Z; Y|C) && (21)\\
&= I(Y; C) + I(Z; Y|C) + I(Z; C|Y) \\
&\quad - I(Z; C) - I(Z; Y|C) && (22)\\
&= I(Y; C) + I(Z; C|Y) - I(Z; C) && (23)\\
&= I(Y; C) - I(Y; C|Z). && (24)
\end{aligned}$$

Since $I(Y; C|Z) \geq 0$, we obtain $\text{Tax}(Y, C) \leq I(Y; C)$.

B Algorithm Details

Algorithm 1 presents the complete CIB fine-tuning procedure. At each training step, we extract representations from both the fine-tuned model and reference model, compute the combined objective (task loss plus contrastive loss), and update the hard negative queue based on computed hardness scores. The queue maintains samples exhibiting both high representation divergence and high task loss, focusing contrastive learning on safety-critical regions.

The hard negative queue update implements a priority queue maintaining the M samples with highest hardness scores. When batch size exceeds queue capacity, we retain only the top-M hardest samples. Otherwise, incoming samples replace queue entries only if their hardness exceeds the current minimum, ensuring the queue focuses on samples where safety-task conflicts are most pronounced.

Algorithm 1 CIB-Guided Fine-tuning

Require: Dataset $\mathcal{D}$, reference model f_{θ_0}, hyperparameters $\{\lambda, \tau, M, K\}$

Initialize LoRA model f_θ and hard negative queue $\mathcal{Q}$
for each batch (X, Y) in $\mathcal{D}$ **do**
  Extract representations: $Z \leftarrow f_\theta(X)$, $Z_{\text{ref}} \leftarrow f_{\theta_0}(X)$
  Sample K hard negatives from $\mathcal{Q}$
  Compute task loss: $\mathcal{L}_{\text{task}} = -\mathbb{E}_{(x,y)}[\log p_\theta(y|x)]$
  Compute contrastive loss: $\mathcal{L}_{\text{safety}}$ (Eq. 15)
  Update model: $\theta \leftarrow \theta - \eta \nabla_\theta(\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{safety}})$
  Compute hardness: $h_i = (1 - \cos(z_i, z_{\text{ref},i})) \cdot \ell_i$
  Update queue $\mathcal{Q}$ with top-M samples by hardness
end for
return Fine-tuned model f_θ

C Experimental Configuration

C.1 Model Architectures and LoRA Settings

We evaluate on Llama-2-7B-Chat (4096 hidden dimensions), Llama-3.1-8B-Instruct (4096 dimensions), Qwen2.5-7B-Instruct (3584 dimensions), and Qwen3-4B-Instruct (2560 dimensions). All models use RoPE positional embeddings and employ post-alignment via RLHF or similar techniques. For LoRA adaptation, we set the rank $r = 32$ and scaling factor $\alpha = 64$ for both single-task and continual-task experiments.

C.2 Training Hyperparameters

We set contrastive weight $\lambda = 5.0$, temperature $\tau = 0.2$, queue size $M = 512$, and sample $K = 64$ negatives per batch. The learning rate is 1×10^{-4} with 50 warmup steps and maximum gradient norm 1.0. We use batch size 16 with gradient accumulation to an effective batch size of 64, training with mixed precision (FP16) on single NVIDIA H100 GPUs (80GB).

C.3 Baseline Configurations

Standard LoRA applies no safety regularization. Vaccine (Huang et al., 2024d) uses adversarial perturbations with $\epsilon = 0.01$ over 3 attack steps during alignment. Safe LoRA (Hsu et al., 2024) performs subspace projection using 512 safety-aligned samples. SaLoRA (Li et al., 2025b) introduces

an orthogonal safety module with rank 16. O-LoRA (Wang et al., 2023) enforces orthogonal subspace constraints. KL regularization applies divergence penalty $\lambda_{KL} = 0.1$ with the reference model.

C.4 Datasets and Evaluation

Training uses Dolly (15k general instructions), GSM8K (7.5k math problems), MedMCQA (20k medical questions), and SQuAD v2 (20k reading comprehension). All data follows the instruction format with explicit instructions and response fields. Safety evaluation uses AdvBench (520 adversarial prompts) and HarmBench (200 diverse safety probes). We compute Harmful Response Rate (HRR) via GPT-4 classification of model outputs. Task performance is measured by cross-entropy loss (Eval) on held-out test sets and zero-shot accuracy on benchmark tasks (ARC-Challenge, GSM8K, ToxiGen, TruthfulQA).

For continual learning experiments, models train sequentially on Dolly → GSM8K → MedMCQA → SQuAD v2, evaluating backward transfer as average accuracy on previously learned tasks after completing the full sequence.

C.5 Prompt Format for Continual Learning

We employ a unified prompt template across all tasks to ensure stability during sequential adaptation:

Unified Prompt Template

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
{instruction}<|im_end|>
<|im_start|>assistant
{response}<|im_end|>
```

This standardization is critical for continual learning stability. Task-specific prompts (e.g., "You are a math tutor" for GSM8K, "You are a medical assistant" for MedMCQA) cause representation drift as models adapt their behavior to different system prompts. The unified format ensures that representation changes reflect genuine task knowledge rather than prompt-induced distribution shifts, enabling more effective safety preservation via our contrastive objective.

We apply this template uniformly to Dolly, GSM8K, MedMCQA, and SQuAD v2, with only the instruction and response content varying per sample. For datasets without explicit instruction-response separation, we construct instructions from available fields (e.g., question text becomes instruction, answer becomes response).

C.6 Evaluation Details

Harmful Response Rate (HRR). 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 (Inan et al., 2023) as an automated safety classifier. For each prompt-response pair (p_i, r_i), Llama-Guard evaluates the conversation against multiple safety categories, including illegal activities, explicit content, hate speech, violence, personal information disclosure, harassment, malicious code, and scams or fraud. A response is classified as harmful if Llama-Guard detects violations in any category. The Harmful Response Rate is computed as:

$$\text{HRR} = \frac{|\{(p_i, r_i) : \textsc{LlamaGuard}(p_i, r_i) = \text{UNSAFE}\}|}{N} \tag{25}$$

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

D Related Work

D.1 Safety Degradation and Defense Methods

Fine-tuning aligned LLMs on benign downstream tasks causes substantial safety degradation (Qi et al., 2023; Peng et al., 2025; Lu et al., 2025). Even without malicious intent, standard fine-tuning on datasets like Alpaca or Dolly compromises safety guardrails through catastrophic forgetting or task-safety conflicts (Huang et al., 2024b,c; Li et al., 2025a). Recent work shows this degradation depends on input formatting and prompt templates, suggesting representation-level vulnerabilities (Huang et al., 2024b).

Existing defenses operate at training time or inference time (Wei et al., 2023). Safe LoRA (Hsu et al., 2024) projects LoRA weights onto safety-aligned subspaces identified from labeled safety data. SaLoRA (Li et al., 2025b) introduces orthogonal safety modules trained separately from task adaptation. Vaccine (Huang et al., 2024d) applies adversarial perturbations during alignment to improve robustness. SafeGrad (Wu et al., 2025) performs gradient surgery to separate task and safety

gradients. Recent representation-based methods like circuit breakers (Zou et al., 2024; Liu et al., 2023; Zhou et al., 2024; Zheng et al.) and contrastive defenses (Casper et al., 2024) target jailbreak robustness by manipulating internal representations.

These approaches share two limitations: they require explicit safety labels or uniform regularization that fails to selectively preserve safety-relevant information, and they lack theoretical guarantees on the safety-performance trade-off. Our work addresses both through information-theoretic analysis.

D.2 Information Bottleneck for LLMs

The Information Bottleneck principle (Tishby et al., 2000) has been applied to various LLM challenges (Butakov et al., 2023). Liu et al. (Liu et al., 2024) use IB to defend against jailbreak attacks by perturbing input prompts, compressing adversarial features while preserving benign semantics. Wang et al. (Wang et al., 2024) apply IB to code generation, reducing memorization of training data through representation compression. Lei et al. (Lei et al., 2024) leverage IB for reasoning tasks, balancing informativeness with generalization across diverse prompts.

No existing work applies the Conditional Information Bottleneck to preserve safety during fine-tuning itself. Prior IB methods target input-level perturbations or task-specific optimization, not the fine-tuning process. We provide the first information-theoretic framework for safety-preserving fine-tuning with formal bounds on the alignment tax.

D.3 Continual Learning for LLMs

Sequential adaptation across multiple tasks exacerbates safety degradation (Qi et al., 2023). O-LoRA (Wang et al., 2023) addresses catastrophic forgetting through orthogonal subspace learning, enabling parameter-efficient continual learning. However, it does not specifically target safety preservation and lacks mechanisms to distinguish safety-relevant from task-specific features, due to recent works focusing more on task performance (Liang et al.; Huang et al., 2024a; Zheng et al., 2025).

Our CIB framework explicitly conditions on aligned reference models throughout sequential adaptation, maintaining safety alignment while adapting to new tasks. The contrastive objective ensures fine-tuned representations retain safety-relevant structure even as models learn task-specific knowledge.

E Detailed Continual Learning Analysis

This appendix provides a comprehensive analysis of the continual learning experiments, examining stage-wise performance patterns across the task sequence: Dolly $\rightarrow$ GSM8K $\rightarrow$ MedMCQA $\rightarrow$ SQuAD v2.

Table 4 presents the complete stage-wise results for all models and methods. Each column group shows performance on the three evaluation tasks (GSM8K, MedMCQA, SQuAD) measured after completing each training stage. This granular view reveals how task knowledge evolves throughout sequential adaptation.

E.1 Stage-wise Forgetting Patterns

We analyze how GSM8K performance evolves across training stages to understand when and where forgetting occurs. Figure 4 visualizes the GSM8K accuracy trajectory for all models and methods throughout the continual learning process.

Key Observations. The GSM8K$\rightarrow$MedMCQA transition (Stage 1$\rightarrow$2) exhibits the most severe forgetting across all models. This is attributable to low task similarity between mathematical reasoning and medical question answering. Notably, **CIB achieves near-zero forgetting on LLaMA-3.1-8B** with only -1.0 point change during Stage 1$\rightarrow$2 and $+0.3$ during Stage 2$\rightarrow$3, while Baseline shows substantial mid-sequence degradation ($+8.4$ points of forgetting). This directly validates Theorem 1's prediction that maximizing $I(Z;C)$ preserves task-relevant structure from the reference model.

E.2 Stability Analysis

Beyond examining forgetting at individual stages, we analyze the overall *stability* of performance across the entire training sequence. We define stability as the standard deviation of task performance measured across all three evaluation stages. Lower values indicate more consistent knowledge retention throughout continual learning.

Figure 6 reveals that **CIB achieves remarkably stable GSM8K performance on LLaMA-3.1-8B** (std=0.42), representing a $10\times$ improvement over Baseline (std=4.27). The trajectory 26.0$\rightarrow$27.0$\rightarrow$26.7 demonstrates that CIB maintains nearly constant mathematical reasoning capability

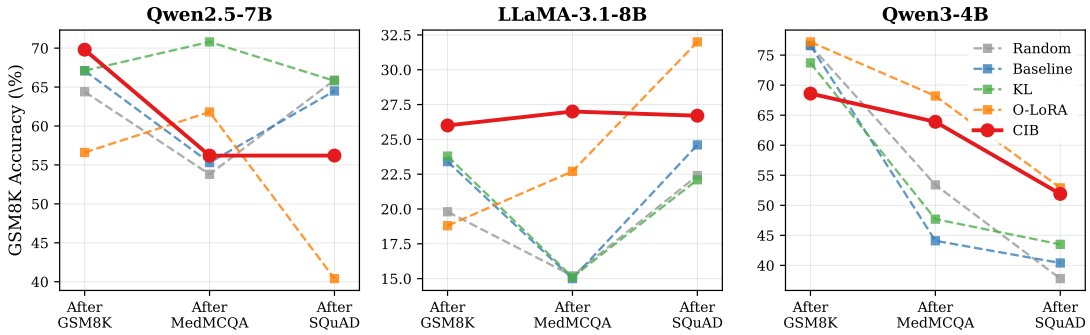

Figure 4: GSM8K performance trajectory during continual learning. Each panel shows one model, with methods distinguished by color and line style. CIB (solid red line) demonstrates remarkable stability on LLaMA-3.1-8B, maintaining performance within ±1 point across all stages (26.0→27.0→26.7), while other methods show substantial fluctuations. On Qwen3-4B, all methods experience severe forgetting, but CIB achieves the most gradual degradation.

		After GSM8K			After MedMCQA			After SQuAD		
Model	**Method**	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	57.3±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
	Contrastive	**69.8**±2.4	55.0±0.4	**52.3**±2.2	56.2±6.4	59.7±0.3	50.7±0.4	56.2±3.4	59.1±0.3	**60.5**±1.4
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±0.4	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
	Contrastive	**26.0**±1.8	58.7±0.4	**52.6**±0.6	**27.0**±1.1	59.7±0.2	51.5±0.7	26.7±2.4	59.1±0.1	68.4±2.3
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
	Contrastive	68.6±0.4	54.8±0.6	50.1±0.0	63.9±4.7	57.9±0.3	50.1±0.0	51.9±3.3	57.5±0.1	56.9±5.3

Table 4: Full continual learning results across training stages. Performance on each task is evaluated after completing GSM8K training (Stage 1), MedMCQA training (Stage 2), and SQuAD training (Stage 3). Values are mean ± std over 3 runs. Bold indicates the best performance among fine-tuned methods for each metric.

despite training on unrelated domains. This stability directly supports our theoretical framework: by maximizing mutual information with the reference model's representations, CIB preserves the representational structure that encodes both safety alignment and general task competence.

On Qwen3-4B, where all methods suffer from high forgetting, CIB still achieves the best stability (std=7.03) compared to Baseline (std=16.26), representing a 2.3× improvement. This suggests that even when absolute forgetting cannot be fully prevented due to architectural factors, CIB provides more graceful degradation. To better understand when forgetting occurs, we decompose the total performance change into stage-wise components. Figure 5 presents a heatmap visualization of GSM8K performance changes during each training transition.

Task-Similarity Effects. The heatmap reveals a clear pattern: forgetting severity correlates with task dissimilarity.

- **Low-similarity transition (GSM8K→MedMCQA):** Mathematical reasoning and medical QA share minimal representational overlap, resulting in 8 to 32 points of forgetting for Baseline across models. CIB uniquely mitigates this on LLaMA-3.1-8B with only −1.0 point change (slight improvement).
- **Higher-similarity transition (MedMCQA→SQuAD):** Both tasks involve

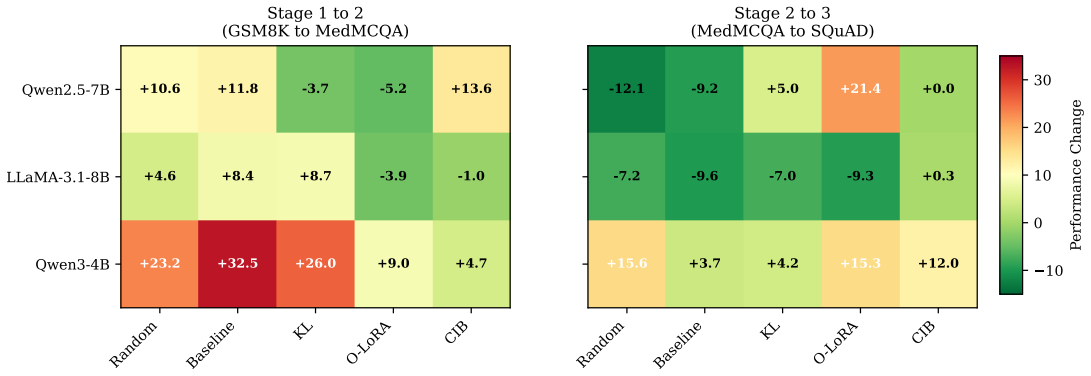

Figure 5: Heatmap of GSM8K performance changes across training stages. Left: Stage 1→2 (GSM8K→MedMCQA training). Right: Stage 2→3 (MedMCQA→SQuAD training). Positive values (red) indicate forgetting; negative values (green) indicate improvement. CIB shows minimal change on LLaMA-3.1-8B across both transitions (-1.0 and $+0.3$), while other methods exhibit high variance.

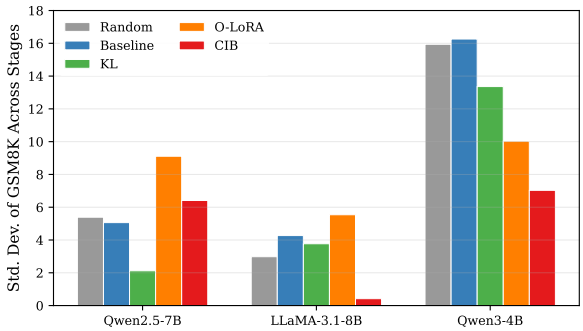

Figure 6: Stability analysis showing standard deviation of GSM8K performance across training stages. Lower values indicate more stable performance. CIB achieves 10× better stability than Baseline on LLaMA-3.1-8B (std=0.42 vs 4.27), demonstrating that maximizing $I(Z;C)$ effectively preserves task knowledge during sequential adaptation.

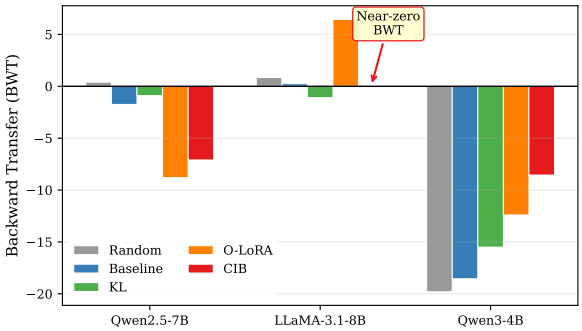

Figure 7: Average Backward Transfer (BWT) across GSM8K and MedMCQA tasks. Negative values indicate forgetting; values closer to zero or positive indicate better knowledge retention. CIB achieves the best BWT on Qwen3-4B (-8.55 vs -18.55 for Baseline) and near-zero BWT on LLaMA-3.1-8B ($+0.05$).

question answering and reading comprehension. All methods show modest MedMCQA forgetting during this stage (<1.5 points), suggesting shared representational requirements between QA tasks.

This pattern connects to our theoretical analysis: the alignment tax bound $I(Y;C)$ is larger when task objectives conflict with safety-relevant structure. For dissimilar tasks, naive fine-tuning may overwrite safety-encoding features to accommodate new task requirements. CIB's contrastive objective explicitly prevents this by maintaining mutual information with the reference model.

E.3 Backward Transfer Analysis

We compute the standard backward transfer (BWT) metric following (Lopez-Paz and Ranzato, 2017):

$$\text{BWT} = \frac{1}{T-1}\sum_{t=1}^{T-1}(R_{T,t} - R_{t,t}) \qquad (26)$$

where $R_{i,j}$ denotes accuracy on task j after training on task i. Negative BWT indicates forgetting, while positive BWT indicates that later training improved performance on earlier tasks.

Figure 7 shows that CIB achieves the best (least negative) BWT on Qwen3-4B (-8.55 vs -18.55 for Baseline) and near-zero BWT on LLaMA-3.1-8B ($+0.05$), indicating effective knowledge retention. The correlation between BWT and safety preservation (Section 5) suggests that methods maintaining stable task performance also better preserve safety alignment.