
Capacity and Redundancy Trade-offs in Multi-Task Learning

Asif Khan¹

¹Harvard Medical School, Boston, USA

Abstract

In multi-task learning (MTL) negative transfer is often considered as an optimization artifact, but it can also be viewed as a consequence of limited shared capacity and weak task redundancy. We investigate this effect through a Capacity–Redundancy (CR) identity that decomposes the sum of per-task predictive informations into joint predictive information that includes label redundancy defined via total correlation, and a residual coupling term that quantifies interference left unresolved by the shared representation. Additionally, we show two key results: (i) a clustering-gap decomposition that gives a necessary and sufficient condition for clustered sharing to outperform global sharing, and (ii) a gradient–TC bridge in a Gaussian multi-task model that formally justifies gradient cosine similarity as a proxy for redundancy ordering. Empirically, we estimate the residual coupling Δ from validation residual correlations, showing that clustered LoRA substantially reduces $\hat{\Delta}$, outperforms size-matched random partitions, and results in statistically significant gains with multi-seed confidence intervals.

1 INTRODUCTION

A single machine learning model trained to optimize multiple prediction tasks simultaneously is useful for improving efficiency and generalization performance [Caruana, 1997, Ruder, 2017]. In a typical setting, MTL is implemented through a shared encoder with task-specific output heads, which allows information to be transferred across tasks while retaining task specialization. Such a training approach has been successfully applied in several areas including computer vision (CV) [Zhang et al., 2014, Misra et al., 2016], natural language processing (NLP) [Collobert

and Weston, 2008], and speech recognition [Deng et al., 2013]. However, despite these successes, there is a key challenge of negative transfer (when training on additional tasks degrades rather than improving performance) [Wang et al., 2019].

Recent developments in large pre-trained language models (LLMs) have renewed interest in MTL in the context of parameter-efficient fine-tuning (PEFT). Methods such as adapters [Houlsby et al., 2019], prefix-tuning [Li and Liang, 2021, Ding et al., 2023], and low-rank adaptation (LoRA) [Hu et al., 2022] make it possible to adapt massive models to new tasks with only a small number of additional parameters. A design choice here is to decide when multiple tasks should share the same lightweight module encouraging cross-task transfer, or should each task be assigned its own module to avoid interference? While previous work have explored both ends as well as various hybrids [He et al., 2021, Karimi Mahabadi et al., 2021, Mahabadi et al., 2021], our focus is on a theoretical framework that can be used to decide when to share and when to specialize model parameters.

Negative transfer occurs when unrelated tasks share a low-rank subspace, whose gradient signals may conflict and cause degradation of performance across all tasks [Yu et al., 2020, Chen et al., 2018]. Giving each task its own private features can avoid this interference but can reduce efficiency. Hybrid approaches such as task clustering, mixture-of-adapters, or routing via task embeddings [Gururangan et al., 2021, Mudrakarta et al., 2018], attempt to balance these trade-offs but lack a unifying theoretical justification.

In this paper, we treat a shared encoder in MTL as a finite-capacity channel, given the budget $I(X; Z_s) \leq C_s$ on the shared latent Z_s . We then prove a CR inequality that shows the total predictive information a single shared latent can provide across tasks is bounded by its capacity plus the label redundancy. Thus, correlations let the same bits be reused across tasks, while weakly related tasks force competition for capacity and make negative transfer unavoidable. We

extend the bound to shared-private representations with per-task budgets and derive conditional variants together with a Bayes-error lower bound that formalize when performance trade-offs cannot be avoided. Finally, we model LoRA as a capacity-constrained channel and show that the CR perspective explains how similarity-based sharing helps when task redundancy is high, whereas low-redundancy task sets require growing effective adapter capacity (rank) or private routes to prevent interference.

In summary, our goal is to develop operational limits of sharing features and use it to prescribe rules for explicit capacity budgets. Our key contributions are: (i) We isolate the slack term $\Delta = \text{TC}(Y^{(1:T)} \mid Z_s)$ and interpret it as interference left after conditioning on the shared representation. (ii) We derive an exact decomposition of the gain of clustered sharing over global sharing into interference reduction minus redundancy loss (Theorem 5.3), which provides a necessary and sufficient condition for when clustering should help. (iii) In a Gaussian multi-task model, we prove that gradient cosine similarity preserves the sign and (under matched noise) the ordering of label correlations, and therefore induces the same clusters as total-correlation-based redundancy (Theorem 4.4). (iv) We show the CR bound is achieved with equality when capacity spans the task subspace and quantify the exact gap under rank constraints via $\text{TC}(Y^{(1:T)} \mid Z_s)$. These results make the intuitive capacity-versus-redundancy tradeoff quantitative: Theorem 5.3 gives an if-and-only-if sharing rule, the Gaussian specialization identifies when the CR bound is tight, and the LoRA rank bound connects adapter capacity to the effective task subspace dimension.

2 RELATED WORK

Multi-task learning. Early work on MTL utilize a shared-backbone as well as task-specific heads to show consistent gain in performance when tasks are related [Caruana, 1997]. In NLP, Collobert and Weston [2008] propose a unified architecture jointly training several sequence tasks. In CV, multi-task feature sharing has been used to exploit structural relatedness between tasks. For example, Zhang et al. [2014] showed improvement in facial landmark detection when jointly trained with auxiliary tasks such as head pose estimation, gender, and smile classification, since these signals share low-level visual features. Similarly, Misra et al. [2016] introduced cross-stitch networks, which explicitly learn linear combinations of feature maps across tasks, allowing the model to adaptively decide how much representation should be shared versus kept task-specific. Beyond deep nets, a large body of regularization approaches formalizes parameter sharing via matrix norms or task relationships [Evgeniou and Pontil, 2004, Ando et al., 2005, Argyriou et al., 2008]. We refer readers to Ruder [2017] for a comprehensive survey of MTL.

Parameter-efficient fine-tuning and adapters. Adapter

layers are used as a light-weight alternative to full fine-tuning by inserting small bottleneck modules into frozen backbones [Houlsby et al., 2019]. Subsequent PEFT methods include prefix/prompt tuning [Li and Liang, 2021]. Ding et al. [2023], He et al. [2021] provide more detailed overviews of PEFT for LLMs. Multi-task PEFT variants allocate a shared adapter or per-task adapters, or combine both through routing, hypernetworks, or mixture strategies [Mahabadi et al., 2021, Karimi Mahabadi et al., 2021, Mu-drakarta et al., 2018, Gururangan et al., 2021]. While empirical heuristics guide which modules to share, principled selection criteria remains limited.

Empirical analyses of task relatedness and negative transfer. Standley et al. [2020] investigate which tasks benefit from being trained together and which cause conflicts, while Zamir et al. [2018] chart the transferability between tasks through a large-scale task graph. Gradient-based analyses of task conflict and mitigation are widely used in practice [Yu et al., 2020, Chen et al., 2018, Chai et al., 2023]. GradNorm [Chen et al., 2018], PCGrad [Yu et al., 2020] treat interference as a gradient conflict to be resolved step-by-step. This works well in practice but it offers no guarantee that interference can be resolved for a given capacity. Broader studies characterize and attempt to avoid negative transfer [Wang et al., 2019]. These methods typically operate at the level of optimization dynamics rather than at the level of distributional limits.

Theoretical generalization and representation sharing. Baxter [2000] formalized task families and bias learning; Evgeniou and Pontil [2004] analyzed kernelized regularization for MTL; and Maurer et al. [2016] show the benefits of learning a shared representation with task-averaged Rademacher complexity bounds. Several other works explicitly partition representations or allocate small task-specific heads to mitigate conflicts [Newell et al., 2020, Wang et al., 2020]. Some other related approaches analyze multi-task trade-offs through multi-objective optimization and controllable Pareto frontiers to show explicit preference control during inference or training [Lin et al., 2020, Momma et al., 2022]. Maurer et al. [2016] bound generalization error and show that sharing helps but their bounds loosen as tasks become dissimilar without explaining where the capacity goes. In contrast, our focus is on a distribution-level converse on the achievable predictive information through a shared latent which is independent of any estimator. Our work is complementary to adaptive-transfer and selective-sharing mechanism using clustered task structure [Kang et al., 2011, Jacob et al., 2008]. Denevi et al. [2022], Chua et al. [2021], and Tian et al. [2025] analyze representation or meta-learning gains through estimator-dependent rates and robustness to similar but nonidentical representations, and Hanneke and Kpotufe [2022] show that multitask benefit cannot be guaranteed without structural assumptions. CR differs by giving a distribution-level converse on predictive

information through a fixed-capacity latent, with the structural condition expressed as capacity versus redundancy.

Multi-task information bottleneck. Multi-task variational information bottleneck (IB) $I(Z; Y) - \beta I(Z; X)$ formulations optimize multi-task IB objectives ($\sum_t I(Z; Y^{(t)})$) directly [Qian et al., 2020], offering algorithmic objectives but not converses [Tishby et al., 2000, Alemi et al., 2016]. Classical results used in our analysis include the independence bound on entropy [Cover, 1999] and its conditional form and the data processing inequality [Cover, 1999], total correlation [Watanabe, 1960] as a redundancy measure, and subset entropy inequalities [Sun, 1975, Madiman and Tetali, 2010]. Our results characterize when sharing is fundamentally limited and when structured sharing (clustering/private capacity) is provably beneficial, independent of a particular training algorithm.

3 CAPACITY-REDUNDANCY INEQUALITY

Let T be the number of tasks with labels $Y^{(1)}, \dots, Y^{(T)}$ and input features X . A shared representation $Z_s = f(X)$ is used across all tasks. Throughout this section, we assume the standard representation setting in which $Y^{(1:T)} - X - Z_s$ forms a Markov chain (equivalently, $I(Z_s; Y^{(1:T)} | X) = 0$), and we impose the shared capacity budget $I(Z_s; X) \leq C_s$. We use total correlation as a measure of label redundancy, $TC(Y^{(1:T)}) = \sum_{t=1}^T H(Y^{(t)}) - H(Y^{(1:T)}) \geq 0$. $TC(Y^{(1:T)}) = 0$ iff the labels are mutually independent. The sum of per-task information carried by a shared bottleneck is always bracketed by the joint information plus redundancy. The main idea requires only the chain rule and data processing, so we focus on implication of bound and for proofs/preliminaries we refer readers to Appendix S1.

Lemma 3.1 (Chain-rule sandwich). *For random variables $(Z, Y^{(1)}, \dots, Y^{(T)})$,*

$$I(Z; Y^{(1:T)}) \leq \sum_{t=1}^T I(Z; Y^{(t)}) \leq I(Z; Y^{(1:T)}) + TC(Y^{(1:T)}). \quad (1)$$

Next, we combine the sandwich with the Markov chain and the capacity budget to define CR inequality.

Theorem 3.2 (CR inequality). *Under the Markov chain and a fixed capacity budget the following bound holds,*

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) \leq C_s + TC(Y^{(1:T)}). \quad (2)$$

and the canonical sandwich Theorem 3.1 holds when $Z = Z_s$,

$$I(Z_s; Y^{(1:T)}) \leq \sum_{t=1}^T I(Z_s; Y^{(t)}) \leq I(Z_s; Y^{(1:T)}) + TC(Y^{(1:T)}). \quad (3)$$

The upper bound in Equation (2) is induced by an exact identity,

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) = I(Z_s; Y^{(1:T)}) + \underbrace{\left(TC(Y^{(1:T)}) - TC(Y^{(1:T)} | Z_s) \right)}_{\text{dependence explained by } Z_s}.$$

which is equivalent to $\sum_{t=1}^T I(Z_s; Y^{(t)}) = I(Z_s; Y^{(1:T)}) + TC(Y^{(1:T)}) - \Delta$, where $\Delta = TC(Y^{(1:T)} | Z_s) \geq 0$.

Here, Δ is the residual multi-task coupling that remains after observing Z_s , $\Delta = 0$ iff the labels become conditionally independent given Z_s . When Δ is large, the right inequality in Equation (3) can be loose, meaning the shared representation has not factorized the task joint distribution, and that additional shared or task-specific capacity (Section 5) is required.

The tightness of Equation (2) depends on conditional independence of tasks given Z_s $\Delta = TC(Y^{(1:T)} | Z_s) \approx 0$, and capacity saturation $I(Z_s; Y^{(1:T)}) \approx I(Z_s; X) \approx C_s$. In a linear-Gaussian setting dependence is explained by Z_s spanning the label-relevant subspace so that cross-task coupling disappears after conditioning, while saturation corresponds to operating at the channel capacity of the bottleneck.

4 CONSEQUENCES AND IMPOSSIBILITY RESULTS

The CR inequality (Theorem 3.2) has a sharp consequence: without redundancy, adding tasks to a fixed-capacity bottleneck must degrade per-task performance as the number of tasks grows unless capacity scales. It further implies that the total per-task information extractable from a shared bottleneck cannot exceed shared capacity plus label redundancy $\sum_{t=1}^T I(Z_s; Y^{(t)}) \leq C_s + TC(Y^{(1:T)})$.

Corollary 4.1 (Average per-task information). *Under the assumptions of Theorem 3.2,*

$$\frac{1}{T} \sum_{t=1}^T I(Z_s; Y^{(t)}) \leq \frac{C_s}{T} + \frac{TC(Y^{(1:T)})}{T}. \quad (4)$$

In particular, there exists at least one task t^ such that $I(Z_s; Y^{(t^*)}) \leq \frac{C_s + TC(Y^{(1:T)})}{T}$.*

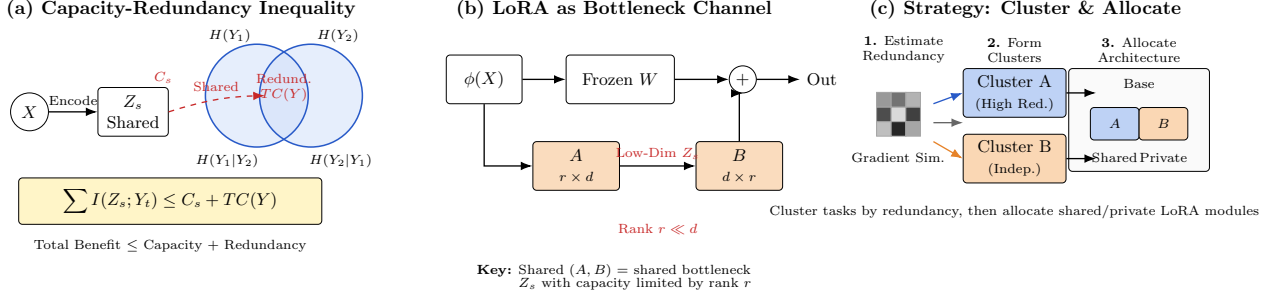


Figure 1: **Capacity-Redundancy Tradeoff in Multi-Task Low-Rank Adaptation.** (a) shows task entropies $H(Y_1)$, $H(Y_2)$ with overlap $TC(Y) = I(Y_1; Y_2)$ representing redundancy. The shared bottleneck Z_s (gray) has capacity C_s . Blue arrows indicate task-specific information $I(Z_s; Y_t)$. (b) rank- r matrices (A, B) add to frozen weights W , where rank $r \ll d$ limits capacity C_s . Sharing (A, B) across tasks creates the shared bottleneck. (c) estimate redundancy via gradient similarity (step 1), cluster tasks by redundancy (step 2), then allocate shared LoRA modules to redundant clusters and private modules to independent tasks (step 3), optimizing the CR bound.

When tasks are weakly dependent, $TC(Y^{(1:T)})$ does not grow with T , and Equation (4) forces the average per-task information to decay like $O(1/T)$ for fixed C_s .

Corollary 4.2 (Linear capacity is necessary for maintaining per-task signal). *Let $TC(Y^{(1:T)}) \leq \tau$ for all T , now for a uniform per-task information level $I(Z_s; Y^{(t)}) \geq b$ for all $t \in [T]$, then necessarily $C_s \geq Tb - \tau$. In particular, for independent tasks ($TC = 0$), maintaining $I(Z_s; Y^{(t)}) \geq b$ for all t forces $C_s \geq Tb$.*

Corollary Theorem 4.2 tells us when negative transfer is unavoidable. If new tasks are added without increasing shared capacity then some tasks must lose predictive signal through the shared bottleneck, unless the new tasks are redundant with existing ones.

Log-loss lower bound. For each task t , let $q_t(\cdot | Z_s)$ denote any predictor and consider the log-loss $\ell_t = -\log q_t(Y^{(t)} | Z_s)$. Then by the Gibbs inequality,

$$\mathbb{E}[\ell_t] \geq H(Y^{(t)} | Z_s) = H(Y^{(t)}) - I(Z_s; Y^{(t)}). \quad (5)$$

Theorem 4.3 (Total log-loss lower bound with residual coupling). *Under the assumptions of Theorem 3.2, for any collection of per-task predictors $\{q_t\}_{t=1}^T$,*

$$\sum_{t=1}^T \mathbb{E}[-\log q_t(Y^{(t)} | Z_s)] \geq \sum_{t=1}^T H(Y^{(t)} | Z_s) = H(Y^{(1:T)} | Z_s) + \Delta, \quad (6)$$

where $\Delta = TC(Y^{(1:T)} | Z_s) \geq 0$. Moreover, using $I(Z_s; Y^{(1:T)}) \leq I(Z_s; X) \leq C_s$,

$$\sum_{t=1}^T \mathbb{E}[-\log q_t(Y^{(t)} | Z_s)] \geq H(Y^{(1:T)}) - C_s + \Delta. \quad (7)$$

The lower bound in Equation (6) implies that $\Delta > 0$ forces extra aggregate log-loss for any set of decoupled per-task predictors $\{q_t(\cdot | Z_s)\}$, unless the shared features Z_s makes tasks conditionally independent. This is why private features (Section 5) are useful as they can reduce Δ by explaining away residual coupling that a shared representation leaves behind.

Appendix Theorem S1.1 gives a Fano-style error lower bound as a corollary, while constants can be loose it still provides a clean impossibility statement directly in terms of misclassification.

When should tasks share? If tasks are redundant (large $TC(Y^{(1:T)})$ or large explained dependence $TC(Y^{(1:T)}) - TC(Y^{(1:T)} | Z_s)$), then sharing features is favorable. If tasks are heterogeneous (small redundancy) and T grows then Theorem 4.2 implies that any globally shared bottleneck must either increase capacity roughly linearly in T or accept degraded per-task signal. Using this observation, we suggest clustered sharing and private capacity allocation as an approach to task heterogeneity.

Gradient cosine similarity as a redundancy proxy. Gradient cosine similarity is widely used to measure task relatedness, but does it track the redundancy that CR predicts. We justify gradient cosine similarity as a proxy for redundancy ordering in a Gaussian model.

Theorem 4.4 (Gradient-TC Bridge (Gaussian multi-task model)). *Let $X \sim \mathcal{N}(0, \Sigma)$ and $Y^{(t)} = w_t^\top X + \varepsilon_t$ with independent $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$. We then define whitened task directions $\tilde{w}_t = \Sigma^{1/2} w_t$ and the label correlation $\rho_{ts} = \frac{w_t^\top \Sigma w_s}{\sqrt{(w_t^\top \Sigma w_t + \sigma_t^2)(w_s^\top \Sigma w_s + \sigma_s^2)}}$. The natural-gradient, equivalently whitened-gradient, cosine similarity at initialization is*

$$G_{ts}^{\text{nat}} = \frac{\tilde{w}_t^\top \tilde{w}_s}{\|\tilde{w}_t\| \|\tilde{w}_s\|} = \frac{w_t^\top \Sigma w_s}{\sqrt{(w_t^\top \Sigma w_t)(w_s^\top \Sigma w_s)}}.$$

Then $\text{sign}(\rho_{ts}) = \text{sign}(G_{ts}^{\text{nat}})$ for all $t \neq s$, and under matched noise levels and equal signal powers $w_t^\top \Sigma w_t$ across tasks, the ordering of $|\rho_{ts}|$ across pairs is the same as the ordering of $|G_{ts}^{\text{nat}}|$.

For the raw-gradient cosine in unwhitened coordinates,

$$G_{ts}^{\text{raw}} = \frac{w_t^\top \Sigma^2 w_s}{\|\Sigma w_t\| \|\Sigma w_s\|},$$

the exact equivalence need not hold under anisotropic Σ ; it coincides with the whitened-gradient form when inputs are whitened or isotropic, with the discrepancy controlled by the conditioning of Σ .

Since for Gaussian labels $\text{TC}(Y^{(1:T)}) = -\frac{1}{2} \log \det R$ where R is the correlation matrix, any clustering rule that depends only on the ranking of pairwise similarities results in the same partition when applied to G^{nat} as when applied to R .

5 SHARED-PRIVATE EXTENSION

In practice, multi-task models often augment shared features with task-private features that can absorb heterogeneity and mitigate interference. A shared-only bound ignores private capacity that most practical architectures already use. We therefore add per task-budgets and derive a corresponding shared-private CR bound. Let $Z_s = f_s(X)$ denote shared features and for each task $t \in [T]$, let $Z_t = f_t(X)$ denote task-private features. The task-specific representation is $Z^{(t)} := (Z_s, Z_t)$. We assume the (test-time) Markov property $Y^{(1:T)} - X - (Z_s, Z_1, \dots, Z_T)$, and impose separated information budgets,

$$I(Z_s; X) \leq C_s, \quad I(Z_t; X | Z_s) \leq C_t \quad \forall t \in [T]. \quad (8)$$

where $I(Z_t; X | Z_s)$ is a conditional budget that measures incremental private capacity beyond what the shared bottleneck already covers. This results in an additive upper bound on the total information extracted from X by the collection $(Z_s, Z_1, \dots, Z_T)$, even when the private feature maps are statistically dependent through X .

Theorem 5.1 (Shared-private capacity-redundancy bound). *Under Markov chain and Equation (8),*

$$\sum_{t=1}^T I(Z^{(t)}; Y^{(t)}) \leq C_s + \sum_{t=1}^T C_t + \text{TC}(Y^{(1:T)}). \quad (9)$$

Theorem 5.1 mirrors the shared-only CR inequality, with the shared capacity C_s replaced by a total budget $C_s + \sum_t C_t$. This shows that private features can reduce the residual task coupling left unexplained by a shared bottleneck. Shared-only slack Δ_s measures how task labels remain coupled after conditioning on the shared representation. When Δ_s is large, the shared representation has not factorized the joint

task structure. In such a setting, global sharing is prone to interference unless C_s grows.

We now use shared-private features to define the post-augmentation residual coupling $\Delta_{sp} = \text{TC}(Y^{(1:T)} | Z_{\text{all}})$, where $Z_{\text{all}} = (Z_s, Z_1, \dots, Z_T)$. Since conditioning cannot increase entropy, $\text{TC}(\cdot | \cdot)$ is monotone in the conditioning set, implying $\Delta_{sp} \leq \Delta_s$. Thus private features can only decrease residual coupling, and doing so is what makes per-task heads easier to fit without forcing the shared bottleneck to encode heterogeneous, task-specific details.

The bound in Equation (9) is tight when the total representation saturates the information budget $I(Z_{\text{all}}; X) \approx C_s + \sum_t C_t$, and the residual dependence is largely removed, $\Delta_{sp} \approx 0$ (labels become approximately conditionally independent given Z_{all}). In contrast, if Δ_s remains large at fixed C_s , then increasing private budgets $\{C_t\}$ is the only way within this framework to reduce Δ_{sp} and avoid the shared-only obstruction.

The shared-private inequality in Equation (9) suggests to share capacity within groups of redundant tasks, and use private capacity to address residual coupling. Later in experiments we validate this by estimating task similarity and comparing global sharing, clustered sharing, and private adapters under matched total rank budgets.

5.1 CLUSTERING AS STRUCTURED PARTIAL SHARING

MTL approaches already cluster tasks and share within clusters. The question is whether this provably helps and by how much. The basic CR bound alone cannot certify when clustering improves over global sharing. We now provide a sharp gain decomposition.

Definition 5.2 (Inter-cluster total correlation). Let $\mathcal{P} = \{S_1, \dots, S_K\}$ be a partition of $[T]$ into disjoint clusters. We define $\text{TC}_{\text{between}}(\mathcal{P}) = \text{TC}(Y^{(1:T)}) - \sum_{k=1}^K \text{TC}(Y^{(S_k)})$, as the redundancy that crosses cluster boundaries.

Theorem 5.3 (Clustering gap decomposition). *Assume the test-time Markov property $Y^{(1:T)} - X - \{Z_s, Z_{S_1}, \dots, Z_{S_K}\}$. Let Z_s^* be a capacity-optimal globally shared encoder under budget $I(Z_s; X) \leq C$. Let $\{Z_{S_k}^*\}_{k=1}^K$ be capacity-optimal per-cluster encoders with budgets $I(Z_{S_k}; X) \leq C_k$ and $\sum_k C_k = C$. We define residual couplings as $\Delta_{\text{sh}} = \text{TC}(Y^{(1:T)} | Z_s^*)$, $\Delta_k = \text{TC}(Y^{(S_k)} | Z_{S_k}^*)$. Then the sum-rate gain of clustering*

over global sharing decomposes as,

$$\sum_{t=1}^T I(Z_{S_k(t)}^*; Y^{(t)}) - \sum_{t=1}^T I(Z_s^*; Y^{(t)}) = \underbrace{\left(\Delta_{\text{sh}} - \sum_{k=1}^K \Delta_k \right)}_{\text{interference reduction} \geq 0} - \underbrace{\text{TC}_{\text{between}}(\mathcal{P})}_{\text{redundancy loss} \geq 0}.$$

Thus clustered sharing is strictly beneficial if and only if the interference reduction exceeds the redundancy loss.

We can use Theorem 5.3 to choose partitions that keep strongly redundant tasks together (small $\text{TC}_{\text{between}}$) while separating antagonistic tasks to reduce residual coupling (large $\Delta_{\text{sh}} - \sum_k \Delta_k$). Fifty et al. [2021] search over task groupings combinatorially, Theorem 5.3 tells us why a particular grouping wins. Here, the interference-reduction term $\Delta_{\text{sh}} - \sum_k \Delta_k$ measures how much residual cross-task coupling is removed when weakly related tasks no longer compete for one shared representation. The redundancy-loss term $\text{TC}_{\text{between}}(\mathcal{P})$ measures useful signal that crosses cluster boundaries and is no longer shared by a single global representation. Thus clustering is most beneficial when the task graph has clear communities with high intra-cluster redundancy and low inter-cluster redundancy.

6 SUBSETS AND SIDE INFORMATION

The global CR bound treats all T tasks as one block, which is wasteful when tasks cluster. In such cases, aggregate CR bounds over all tasks may be too coarse to reflect finer-grained task relationships. Subset bounds exploit this structure.

Subset bounds. Let $S \subseteq [T]$ be any subset of tasks and denote $Y^S = \{Y^{(t)} : t \in S\}$. Then the subset total correlation is defined as $\text{TC}(Y^S) = \sum_{t \in S} H(Y^{(t)}) - H(Y^S)$. Applying Theorem 3.1 and Theorem 3.2 to the subcollection results in an immediate tightening.

Corollary 6.1 (Subset CR inequality). *Under the assumptions of Theorem 3.2, for any $S \subseteq [T]$,*

$$\sum_{t \in S} I(Z_s; Y^{(t)}) \leq C_s + \text{TC}(Y^S). \quad (10)$$

Moreover, the canonical sandwich holds on S ,

$$I(Z_s; Y^S) \leq \sum_{t \in S} I(Z_s; Y^{(t)}) \leq I(Z_s; Y^S) + \text{TC}(Y^S). \quad (11)$$

The subset bound is usually tighter than Equation (2) because redundancy is not uniform, tasks often form clusters with high intra-cluster dependence but low inter-cluster dependence. This refinement suggests that shared capacity

should be allocated relative to $\text{TC}(Y^S)$ within each cluster, with respect to the global quantity $\text{TC}(Y^{(1:T)})$.

Side information and conditional CR. Let W denote side information available when constructing or analyzing the representation. The conditional total correlation is defined as $\text{TC}(Y^{(1:T)} | W) = \sum_{t=1}^T H(Y^{(t)} | W) - H(Y^{(1:T)} | W) \geq 0$.

Theorem 6.2 (Conditional CR inequality). *Under a conditional Markov property $Y^{(1:T)} - X - Z_s$ given W , and a shared capacity budget $I(Z_s; X | W) \leq C_s(W)$,*

$$\sum_{t=1}^T I(Z_s; Y^{(t)} | W) \leq C_s(W) + \text{TC}(Y^{(1:T)} | W). \quad (12)$$

More generally, for any subset $S \subseteq [T]$,

$$\sum_{t \in S} I(Z_s; Y^{(t)} | W) \leq C_s(W) + \text{TC}(Y^S | W). \quad (13)$$

Residual coupling with side information. The exact decomposition also holds conditionally $\sum_{t=1}^T I(Z_s; Y^{(t)} | W) = I(Z_s; Y^{(1:T)} | W) + \left(\text{TC}(Y^{(1:T)} | W) - \text{TC}(Y^{(1:T)} | Z_s, W) \right)$ and the conditional slack $\Delta(W) = \text{TC}(Y^{(1:T)} | Z_s, W) \geq 0$ again measures residual task coupling after observing Z_s , now at a fixed value of side information W . During training, the representation parameters depend on labels, so the unconditional Markov chain $Y^{(1:T)} - X - Z_s$ need not hold. However, at evaluation time weights are fixed making it possible to use conditional Markov property, DPI and CR arguments at test time. This is crucial as it separates learning dynamics from the representational limit that are the main focus of this work.

If W includes task identity or context, then $\text{TC}(Y^{(1:T)} | W)$ can be much smaller than $\text{TC}(Y^{(1:T)})$, since conditioning can decouple labels across tasks. In such a setting, using conditional adapters or MoE routing increase effective performance by reducing conditional redundancy terms or allowing $C_s(W)$ to vary with W . The conditional CR bound Equation (12) serves as a unified lens for these designs. It also covers task-dependent input distributions by taking W to include task identity or context.

Implications for clustered sharing. Subset and conditional refinements together suggest to first estimate task similarity (as a proxy for dependence), cluster tasks into subsets $\{S_k\}$, and allocate shared capacity per cluster. Formally, applying Equation (10) within each cluster gives $\sum_{t \in S_k} I(Z_s^{(k)}; Y^{(t)}) \leq C_s^{(k)} + \text{TC}(Y^{S_k})$ so at fixed total budget $\sum_k C_s^{(k)}$, clustered sharing targets the local redundancy structure rather than paying for global heterogeneity. We refer readers to Appendix Section S2 for a Gaussian model where both shared capacity and redundancy can be computed in closed form.

	r=4	r=8	r=16
Shared LoRA	0.763	0.773	0.755
Private LoRA	0.796	0.795	0.797
Clustered LoRA	0.812	0.817	0.806

Table 1: **Mean validation accuracy** on SST2/MRPC/RTE for budget-valid total ranks. Clustered sharing consistently outperforms global sharing under matched rank budgets.

7 LORA FINE-TUNING

Next, we introduce a capacity proxy to link adapter rank to an effective shared budget C_s that results in a testable predictions about when globally shared low-rank updates saturate and when clustered/private updates are necessary.

LoRA as a low-rank information channel. Consider a frozen backbone with a hidden representation $\phi(X) \in \mathbb{R}^d$. A LoRA update replaces a linear map $W \in \mathbb{R}^{o \times d}$ by

$$W \mapsto W + \Delta W, \quad \Delta W = BA, \quad A \in \mathbb{R}^{r \times d}, \quad B \in \mathbb{R}^{o \times r}, \quad (14)$$

where $r \ll \min\{o, d\}$ is the LoRA rank. We view LoRA branch as a feature channel that compresses the input into a low-dimensional signal $Z_s = A\phi(X) \in \mathbb{R}^r$, which is then mixed into the output via B . In a multi task setup, this implies a structural constraint that tasks sharing the same LoRA factors share the same bottleneck Z_s . Thus, the shared representation capacity is controlled by the rank r and the scale of A relative to the feature distribution $\phi(X)$. This is consistent with the intrinsic-dimensionality perspective [Aghajanyan et al., 2021].

CR predictions for LoRA design. Let us consider T tasks fine-tuned by sharing the same LoRA factors (A, B) (and thus the same bottleneck Z_s). Applying Theorem 3.2 with shared budget $C_s \approx I(Z_s; \phi(X))$ the capacity can be upper bound via Equation (S3.32). If tasks are heterogeneous so that $TC(Y^{1:T})$ is small, then CR prediction implies that the sum of task-relevant information is controlled primarily by $I(Z_s; \phi(X))$, which grows slowly with r . Consequently, performance under a single globally shared LoRA should saturate quickly as rank increases, and adding tasks without increasing rank induces interference (Section 4).

To obtain a per-task information level $I(Z_s; Y^{(t)}) \gtrsim b$ across many weakly dependent tasks, Corollary 4.2 tells us that the total shared information $I(Z_s; \phi(X))$ must scale approximately like Tb (modulo redundancy). Under our rank-controlled bound, LoRA rank must grow roughly linearly in the number of distinct task directions or clusters. If tasks form clusters $\{S_k\}$ with high within-cluster dependence but low across-cluster dependence, then we can use the subset CR bound (Theorem 6.1), sharing LoRA within

each cluster and allocating rank per cluster,

$$\sum_{t \in S_k} I(Z_s^{(k)}; Y^{(t)}) \leq C_s^{(k)} + TC(Y^{S_k}), \quad \sum_k C_s^{(k)} \leq C_{\text{tot}}.$$

Estimating residual coupling Δ from validation residuals. We use validation data $\{(x_i, y_i^{(1:T)})\}_{i=1}^n$ to compute per-task residuals $e_t^{(i)} = y_t^{(i)} - \hat{y}_t^{(i)}$. Let $\hat{R}_e$ be the sample correlation matrix of $(e_1, \dots, e_T)$. We approximate $\hat{\Delta} = -\frac{1}{2} \log \det(\hat{R}_e + \epsilon I) \approx TC(Y^{(1:T)} | Z_s)$, and report it across different settings. In the NLP experiments we use $\epsilon = 10^{-5}$; the synthetic sensitivity sweep below shows this value is in the near-unbiased regime for our validation sample sizes, while larger ϵ persistently underestimates the conditional TC.

8 EMPIRICAL VALIDATION

8.1 VALIDATION ON SYNTHETIC DATA.

CR inequality for Gaussian case. We first verify the CR inequality in a linear-Gaussian multitask model where all terms are available in closed form, and show that the bound is tight once the representation spans the task subspace. In particular, when the encoder row-space contains the span of task directions, the residual coupling term vanishes, while under rank constraints the CR slack equals the conditional total correlation $TC(Y^{1:T} | Z)$. Full derivations and controlled sweeps across redundancy, encoder alignment, and rank are provided in Appendix S7.

Validation for $\hat{\Delta}$. We decouple estimator behavior from representation learning using a synthetic Gaussian multitask model where the target conditional TC quantity is known in closed form. We then estimate $\hat{\Delta}$ from N i.i.d. samples and study both absolute error and signed bias across ridge parameters ϵ . Figure 2 shows that small ϵ in a near-unbiased estimates as N increases, while larger ϵ induces a persistent negative bias that does not vanish with sample size. At the largest sample size ($N = 16384$), Table 2 quantifies this effect $\epsilon \in \{10^{-6}, 10^{-5}\}$ achieves small error, whereas $\epsilon \geq 10^{-4}$ results in an underestimation. This shows that the ϵ acts as a regularization term for a log-determinant inverse problem, and using small ϵ in downstream evaluations. This sensitivity analysis directly motivates the choice $\epsilon = 10^{-5}$ used for the LoRA residual-coupling estimates.

8.2 CR ALIGNED PEFT.

We now evaluate applicability of CR in a real setting by training multi-task PEFT with a frozen bert-base-uncased encoder and LoRA adapters applied to attention query and value projections.

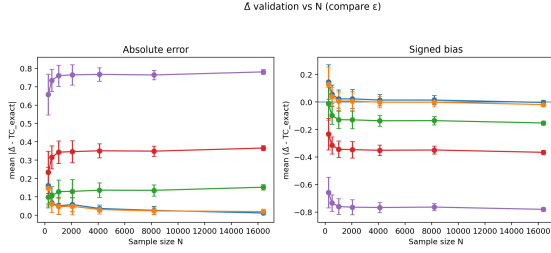


Figure 2: Mean absolute error $|\hat{\Delta} - TC_{\text{exact}}|$ vs sample size N on the left. Signed bias $\hat{\Delta} - TC_{\text{exact}}$ vs N on the right. Small ε converges toward the ground truth as N increases, while larger ε results in a persistent negative bias.

Table 2: Gaussian setting: error of $\hat{\Delta}$ at the largest sample size ($N = 16384$, 10 seeds).

ε	mean $ \hat{\Delta} - TC $	std	mean signed	std
1e-06	0.0123	0.0089	-0.0037	0.0153
1e-05	0.0201	0.0138	-0.0191	0.0152
1e-04	0.1528	0.0148	-0.1528	0.0148
3e-04	0.3660	0.0143	-0.3660	0.0143
1e-03	0.7807	0.0138	-0.7807	0.0138

GoEmotions Dataset. We use GoEmotions [Demszky et al., 2020] dataset which consists of shared inputs and multiple binary labels. We treat each label as a task sharing the same input text and select twelve most frequent GoEmotions labels (fixed across runs) and evaluate each task using AUROC on the validation split and report the macro-average AUROC across the chosen tasks.

We analyze two finetuning setting under a fixed LoRA rank budget $r_{\text{tot}} \in \{4, 8, 12, 16, 24, 32\}$: using one shared adapter across all tasks, and cluster tasks into K clusters based on similarity and use K adapters. For clustered routing, the clustering is done once using gradient-similarity probes (details in Appendix S6). We also track $\hat{\Delta}$ as a diagnostic of residual dependence. Figure 3 shows that similarity-based routing substantially improves performance under small budgets. From $r_{\text{tot}} = 8$ onward, clustered consistently outperforms shared in average AUCROC (Table 3). Moreover, the clustered routing results in a systematically lower $\hat{\Delta}$ at matched rank budgets (Figure 4) which shows routing reduces residual interference in the learned predictor.

GLUE8 benchmark. We evaluate on eight GLUE tasks (SST-2, MRPC, RTE, QNLI, QQP, MNLI, CoLA, STS-B) using task metrics and report the average score across tasks. (Appendix S6). Figure 5 shows that clustered and private adapters provide modest but consistent gains over naive sharing for budgets where routing is feasible ($\text{rank} \geq 8$ in our setup). Unlike GoEmotions, the improvements on GLUE are smaller in magnitude, which can be attributed

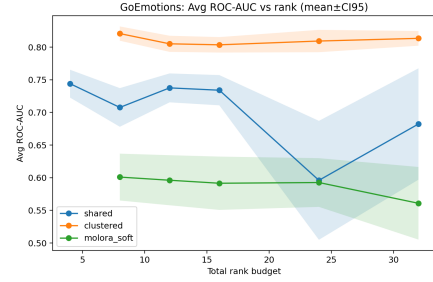


Figure 3: Macro-average AUROC on the validation split as a function of total LoRA rank budget r_{tot} . Clustered routing consistently improves over naive sharing from $r_{\text{tot}} = 8$ onward. Shaded regions denote CI.



Figure 4: $\hat{\Delta}$ computed from validation residuals vs total rank budget. Clustered routing results in a systematically smaller $\hat{\Delta}$ than sharing at matched budgets ($\text{rank} \geq 8$), showing reduced residual dependence under similarity-guided finetuning.

to heterogeneous task formats and limited sharing opportunities once syntax and sentiment tasks are isolated by similarity. This asymmetry is consistent with the CR prediction: GoEmotions labels are subcategories of a shared emotion taxonomy with higher cross-label redundancy, whereas GLUE8 combines heterogeneous task formats with weaker cross-task structure, so clustering has less redundant signal to reuse. Nevertheless, the results show similarity-guided finetuning improves transfer under a parameter budget.

Overall, our empirical results across GoEmotions and GLUE8 demonstrate that routing based on gradient similarity improves performance under fixed adapter capacity. So when tasks compete for limited shared capacity a naive sharing can amplify interference while specialization restores effective capacity aligned with our CR perspective.

On GoEmotions, clustered routing reduces $\hat{\Delta}$ alongside improving AUROC (Figure 4). The Gaussian experiment clarifies how $\hat{\Delta}$ acts like a regularization, increasing N reduces variance, while ε controls bias (Figure 2). Thus, $\hat{\Delta}$ can be used as a compact diagnostic of residual dependence in multi-task settings.

Table 3: Average AUROC (mean $\pm$ CI95) vs total rank budget.

Rank	shared	clustered
4	0.744 $\pm$ 0.021	—
8	0.708 $\pm$ 0.030	0.821 $\pm$ 0.011
12	0.738 $\pm$ 0.022	0.805 $\pm$ 0.012
16	0.734 $\pm$ 0.023	0.804 $\pm$ 0.012
24	0.596 $\pm$ 0.091	0.809 $\pm$ 0.017
32	0.682 $\pm$ 0.085	0.813 $\pm$ 0.011

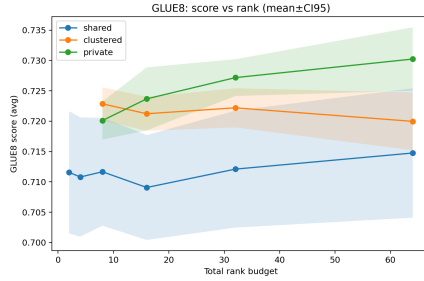


Figure 5: Shared is evaluated at all budgets, clustered/private are feasible at rank ≥ 8 due to the number of adapters/tasks. We observe consistent improvement over naive sharing for the feasible budgets.

9 CONCLUSION

We introduced the CR inequality, a distribution-level converse that upper-bounds the total predictive information extractable from a shared latent by its information capacity plus the redundancy among tasks. This result explains negative transfer as an unavoidable consequence of low redundancy and finite capacity, and provides a principled approach for the design of shared–private architectures, including LoRA-based fine-tuning schemes. Furthermore, we demonstrate a similarity-guided routing approach in PEFT improves performance on multitask under strict capacity budgets. Overall, our results suggest that under limited adapter capacity sharing parameters based on gradient based task similarity provides a practical basis for budget-aware specialization. Because TC measures redundancy, not synergy, CR does not capture complementary information across tasks. Extensions via partial information decomposition or multivariate interaction measures (e.g., *O*-information [Rosas et al., 2019]) are possible future directions to go beyond redundancy. Moreover, extending the analysis to dynamic training settings may explain how optimization implicitly allocates capacity across tasks.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of

Table 4: GLUE8: average score vs total rank budget (mean $\pm$ CI95) using official task metrics.

rank	regime	n	score mean	score ci95
2	shared	10	0.712	0.010
2	clustered	0	—	—
2	private	0	—	—
4	shared	10	0.711	0.010
4	clustered	0	—	—
4	private	0	—	—
8	shared	10	0.712	0.009
8	clustered	10	0.723	0.003
8	private	10	0.720	0.003
16	shared	10	0.709	0.009
16	clustered	10	0.721	0.003
16	private	10	0.724	0.005
32	shared	10	0.712	0.010
32	clustered	10	0.722	0.003
32	private	10	0.727	0.003
64	shared	7	0.715	0.011
64	clustered	7	0.720	0.005
64	private	6	0.730	0.005

language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.568. URL <https://aclanthology.org/2021.acl-long.568/>.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Heyan Chai, Jinhao Cui, Ye Wang, Min Zhang, Binxing Fang, and Qing Liao. Improving gradient trade-offs between tasks in multi-task text classification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki,

- editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2565–2579, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.144. URL <https://aclanthology.org/2023.acl-long.144/>.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.
- Kurtland Chua, Qi Lei, and Jason D. Lee. How fine-tuning allows for effective meta-learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.
- Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL <https://aclanthology.org/2020.acl-main.372/>.
- Giulia Denevi, Massimiliano Pontil, and Carlo Ciliberto. Conditional meta-learning of linear representations. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. *ICASSP*, pages 8599–8603, 2013.
- Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235, 2023.
- Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.
- Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.
- Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*, 2021.
- Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6), 2022.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bryan Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799, 2019.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems*, volume 21, 2008.
- Zhuoli Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 521–528, 2011.
- Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in neural information processing systems*, 34:1022–1035, 2021.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL <https://aclanthology.org/2021.acl-long.353/>.
- Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhu, and James Kwok. Controllable pareto multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20337–20348, 2020. URL <https://proceedings.neurips.cc>.

- Mokshay Madiman and Prasad Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Transactions on Information Theory*, 56(6):2699–2713, 2010.
- Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.
- Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016.
- Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective / multi-task learning framework induced by pareto stationarity. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15895–15907. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/momma22a.html>.
- Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, and Andrew Howard. K for the price of 1: Parameter-efficient multi-task and transfer learning. *arXiv preprint arXiv:1810.10703*, 2018.
- Alejandro Newell, Lu Jiang, Chong Wang, Li-Jia Li, and Jia Deng. Feature partitioning for efficient multi-task architectures, 2020. URL <https://openreview.net/forum?id=BleoyAVFwH>.
- Weizhu Qian, Bowei Chen, Yichao Zhang, Guanghui Wen, and Franck Gechter. Multi-task variational information bottleneck. *arXiv preprint arXiv:2007.00339*, 2020.
- Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305, 2019.
- Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.
- TH Sun. Linear dependence structure of the entropy space. *Inf. Control*, 29(4):337–368, 1975.
- Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *Journal of Machine Learning Research*, 2025.
- Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- Yuyan Wang, Zhe Zhao, Bo Dai, Christopher Fifty, Dong Lin, Lichan Hong, and Ed H Chi. Small towers make big differences. *arXiv preprint arXiv:2008.05808*, 2020.
- Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.
- Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.
- Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.
- Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

SUPPLEMENTARY MATERIAL

S1 NOTATION AND PRELIMINARIES

Symbol	Meaning
X	input variable
$Y_{1:T} = (Y_1, \dots, Y_T)$	task labels
Z	learned representation (shared or composite)
Z_s	shared representation
Z_t	private representation for task t
W	parameterized weight matrix
$H(\cdot)$	Shannon entropy
$I(A; B C)$	conditional mutual information
$\text{TC}(Y_{1:T} U)$	conditional total correlation
C_s, C_t	capacity budgets (information or proxy budget)

Total correlation. For label sequence $Y_{1:T}$, the TC and conditional TC terms are defined as,

$$\text{TC}(Y_{1:T}) = \sum_{t=1}^T H(Y_t) - H(Y_{1:T}), \quad \text{TC}(Y_{1:T} | U) = \sum_{t=1}^T H(Y_t | U) - H(Y_{1:T} | U). \quad (\text{S1.15})$$

where $\text{TC}(Y_{1:T} | U) \geq 0$ with equality iff $Y_1, \dots, Y_T$ are conditionally independent given U .

Identity underlying CR. For any representation Z and side information W ,

$$\sum_{t=1}^T I(Z; Y_t | W) = I(Z; Y_{1:T} | W) + \text{TC}(Y_{1:T} | W) - \text{TC}(Y_{1:T} | Z, W). \quad (\text{S1.16})$$

Corollary S1.1 (Per-task error lower bound via Fano). *Assume each $Y^{(t)}$ takes values in a finite alphabet of size $|\mathcal{Y}_t| \geq 2$ and define the Bayes error given Z_s , $P_{e,t}^* = \inf_{\hat{y}_t(Z_s)} \mathbb{P}[\hat{y}_t(Z_s) \neq Y^{(t)}]$. Then*

$$P_{e,t}^* \geq \frac{H(Y^{(t)} | Z_s) - 1}{\log |\mathcal{Y}_t|} = \frac{H(Y^{(t)}) - I(Z_s; Y^{(t)}) - 1}{\log |\mathcal{Y}_t|}. \quad (\text{S1.17})$$

There exists a task t^* with

$$P_{e,t^*}^* \gtrsim \frac{H(Y^{(t^*)}) - \frac{C_s + \text{TC}(Y^{(1:T)})}{T}}{\log |\mathcal{Y}_{t^*}|}. \quad (\text{S1.18})$$

S2 GAUSSIAN SPECIALIZATION AND TIGHTNESS

A linear-Gaussian model offers an interpretable setting where we can evaluate the CR bounds exactly. Additionally, it shows exact conditions such as span alignment, orthogonality, and residual coupling under which these bounds become tight.

Linear-Gaussian setup Let $X \sim \mathcal{N}(0, \Sigma_X) \in \mathbb{R}^d$, we assume tasks are generated as noisy linear measurements,

$$Y^{(t)} = u_t^\top X + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \quad (\text{S2.19})$$

where $\varepsilon_{1:T}$ independent of X , we write $Y^{(1:T)} = U^\top X + \varepsilon$ where $U = [u_1, \dots, u_T] \in \mathbb{R}^{d \times T}$ and $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$ with $\Sigma_\varepsilon = \text{diag}(\sigma_1^2, \dots, \sigma_T^2)$. A shared representation is produced by a linear map with an additive Gaussian noise:

$$Z_s = AX + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_z^2 I_m), \quad \eta \perp (X, \varepsilon), \quad (\text{S2.20})$$

where $A \in \mathbb{R}^{m \times d}$ is a design matrix, and σ_z^2 controls the bottleneck noise.

S2.1 CLOSED-FORM CAPACITY AND REDUNDANCY

For the Gaussian channel (S2.20),

$$I(Z_s; X) = \frac{1}{2} \log \det \left(I_m + \frac{1}{\sigma_z^2} A \Sigma_X A^\top \right). \quad (\text{S2.21})$$

The above expression makes the capacity dependence on the spectrum of $A \Sigma_X^{1/2}$ explicit. At fixed output dimension m , capacity is governed by the nonzero singular values and saturates when $A \Sigma_X A^\top$ becomes low-rank or small-norm.

Label redundancy. Since $Y^{(1:T)}$ is jointly Gaussian with covariance $\Sigma_Y = U^\top \Sigma_X U + \Sigma_\epsilon$, the total correlation reduces to a log-determinant ratio,

$$TC(Y^{(1:T)}) = \frac{1}{2} \log \frac{\prod_{t=1}^T \Sigma_{Y,tt}}{\det(\Sigma_Y)}. \quad (\text{S2.22})$$

$TC(Y^{(1:T)})$ is large when task outputs are strongly correlated (e.g., when $\{u_t\}$ have large overlap in the Σ_X -geometry), and vanishes when Σ_Y is diagonal (independent tasks).

CR bound is tight in a Gaussian specialization under capacity saturation $I(Z_s; Y^{(1:T)}) \approx I(Z_s; X)$, and explained dependence $\Delta = TC(Y^{(1:T)} | Z_s) \approx 0$ (labels become nearly conditionally independent given Z_s).

When sharing is optimal. In the Gaussian setting $I(Z_s; Y^{(1:T)})$ depends on how well the row space of A captures the subspace spanned by $\{u_t\}$. Let the task subspace be $\mathcal{U} := \text{span}\{u_1, \dots, u_T\} \subseteq \mathbb{R}^d$ in the Σ_X -geometry. If the row space of A contains $\mathcal{U}$ (equivalently, AX preserves all task-relevant directions), then Z_s is a sufficient statistic for $U^\top X$ up to noise, and the joint information $I(Z_s; Y^{(1:T)})$ approaches $I(Z_s; X)$ as $\sigma_z^2 \rightarrow 0$ under a fixed power constraint on A . Thus, conditioning on Z_s removes most of the cross-task dependence induced by the shared latent X , so $\Delta \approx 0$ when the representation is informative enough about $U^\top X$. This results in a near-tightness of CR,

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) \approx I(Z_s; X) + TC(Y^{(1:T)}). \quad (\text{S2.23})$$

When sharing must fail at fixed rank. Suppose task directions are nearly orthogonal in the Σ_X inner product, so that $U^\top \Sigma_X U$ is close to diagonal. Then $TC(Y^{(1:T)})$ is small and there is little redundancy to reuse across tasks. Additionally, the representation rank is limited then Z_s cannot preserve all task directions simultaneously. In this scenario, $I(Z_s; Y^{(t)})$ must be small for many tasks and the sum information is bounded by capacity alone,

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) \lesssim I(Z_s; X) \leq C_s, \quad (\text{S2.24})$$

when $TC(Y^{(1:T)}) \approx 0$. This shows negative transfer under global sharing adding heterogeneous tasks forces interference unless the shared capacity scales with the number of distinct task directions.

Proposition S2.1 (Near-tightness under span capture). *Assume $\Sigma_X = I_d$ and $\sigma_t^2 = \sigma^2$ for all t . Let $\mathcal{U} = \text{span}\{u_1, \dots, u_T\}$ have dimension k and suppose A has rank at least k and row space containing $\mathcal{U}$. Under a fixed power constraint $\text{tr}(AA^\top) \leq P$, there exists a choice of A and σ_z^2 such that*

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) \geq I(Z_s; X) + TC(Y^{(1:T)}) - \epsilon. \quad (\text{S2.25})$$

S3 FULL PROOFS

S3.1 CR INEQUALITY AND SUBSET REGION

Theorem S3.1 (CR bound). *Assume the Markov chain $Y_{1:T} - X - Z_s$ holds given W and the fixed capacity constraint $I(X; Z_s | W) \leq C_s$. Then*

$$\sum_{t=1}^T I(Z_s; Y_t | W) \leq C_s + TC(Y_{1:T} | W). \quad (\text{S3.26})$$

For any subset $S \subseteq [T]$ we can write,

$$\sum_{t \in S} I(Z_s; Y_t | W) \leq C_s + \text{TC}(Y_S | W). \quad (\text{S3.27})$$

Proof. Using CR identity with $Z = Z_s$ and non-negativity of TC,

$$\sum_{t=1}^T I(Z_s; Y_t | W) \leq I(Z_s; Y_{1:T} | W) + \text{TC}(Y_{1:T} | W).$$

By data-processing under $Y_{1:T} - X - Z_s$ given W , $I(Z_s; Y_{1:T} | W) \leq I(Z_s; X | W) \leq C_s$. The subset statement repeats the same argument for Y_S . $\square$

S3.2 SHARED-PRIVATE EXTENSION

Theorem S3.2 (Shared-private CR bound). *Assume $Y_{1:T} - X - (Z_s, Z_{1:T})$ is a Markov chain given W and $I(X; Z_s | W) \leq C_s$, $I(X; Z_t | W) \leq C_t$ for each t . Then,*

$$\sum_{t=1}^T I((Z_s, Z_t); Y_t | W) \leq C_s + \sum_{t=1}^T C_t + \text{TC}(Y_{1:T} | W). \quad (\text{S3.28})$$

Proof. For each task t , $I((Z_s, Z_t); Y_t | W) \leq I((Z_s, Z_{1:T}); Y_t | W)$. Summing over t and applying CR identity with $Z = (Z_s, Z_{1:T})$ gives,

$$\sum_{t=1}^T I((Z_s, Z_{1:T}); Y_t | W) \leq I((Z_s, Z_{1:T}); Y_{1:T} | W) + \text{TC}(Y_{1:T} | W).$$

Using data-processing, $I((Z_s, Z_{1:T}); Y_{1:T} | W) \leq I((Z_s, Z_{1:T}); X | W)$ with chain rule and non-negativity,

$$I((Z_s, Z_{1:T}); X | W) \leq I(Z_s; X | W) + \sum_{t=1}^T I(Z_t; X | W) \leq C_s + \sum_{t=1}^T C_t.$$

$\square$

S3.3 LORA CAPACITY BOUND (GAUSSIAN-NOISE CHANNEL)

Theorem S3.3 (Linear-Gaussian channel upper bound). *Let U be any random vector with finite covariance $\Sigma_U = \text{Cov}(U | W)$. Let $Z = AU + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I)$ independent of U given W . Then*

$$I(U; Z | W) \leq \frac{1}{2} \log \det(I + \sigma^{-2} A \Sigma_U A^\top). \quad (\text{S3.29})$$

If additionally $\text{rank}(A) \leq r$ and $\text{tr}(A \Sigma_U A^\top) \leq \kappa$, then

$$I(U; Z | W) \leq \frac{r}{2} \log \left(1 + \frac{\kappa}{r \sigma^2} \right). \quad (\text{S3.30})$$

Proof. Conditioning on W , we can write $\Sigma_Z = A \Sigma_U A^\top + \sigma^2 I$. Since $Z|U$ is a Gaussian noise, $h(Z | U) = h(\xi)$. By maximum-entropy $h(Z) \leq \frac{1}{2} \log((2\pi e)^m \det(\Sigma_Z))$. Thus $I(U; Z) = h(Z) - h(Z | U) \leq \frac{1}{2} \log \frac{\det(\Sigma_Z)}{\det(\sigma^2 I)}$. For the rank/trace bound, apply AM-GM on nonzero eigenvalues of $M = \sigma^{-2} A \Sigma_U A^\top \succeq 0$: $\log \det(I + M) \leq r \log(1 + \text{tr}(M)/r)$. $\square$

A Gaussian surrogate for LoRA capacity. Consider a Gaussian/noisy-channel surrogate at inference time,

$$Z_s = A\phi(X) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_r), \quad (\text{S3.31})$$

where η is a noise used to make mutual information finite for deterministic encoders. Let $\Sigma_\phi := \text{Cov}(\phi(X))$. Then, by the Gaussian channel formula,

$$I(Z_s; \phi(X)) = \frac{1}{2} \log \det \left(I_r + \frac{1}{\sigma^2} A \Sigma_\phi A^\top \right). \quad (\text{S3.32})$$

where $\phi(X)$ is a deterministic function of X given frozen weights. Now using DPI $I(Z_s; X) \geq I(Z_s; \phi(X))$ and $I(Z_s; Y^{(1:T)}) \leq I(Z_s; X)$, which implies (S3.32) is a conservative, architecture-linked proxy for the shared budget.

Rank- and power-controlled upper bound. Assume a power constraint $\text{tr}(AA^\top) \leq \kappa$ and denote by $\lambda_1(\Sigma_\phi) \geq \dots \geq \lambda_d(\Sigma_\phi)$ the eigenvalues of Σ_ϕ . By eigenvalue majorization and log-det bounds, $I(Z_s; \phi(X)) \leq \frac{r}{2} \log \left(1 + \frac{\kappa}{r\sigma^2} \lambda_1(\Sigma_\phi) \right) \leq \frac{1}{2} \sum_{i=1}^r \log \left(1 + \frac{\kappa}{\sigma^2} \lambda_i(\Sigma_\phi) \right)$.

This shows that effective shared capacity grows at most linearly with rank r , and it saturates if Σ_ϕ is low-rank or if κ is small.

S4 PROOF OF THEOREM 5.3

Proof of Theorem 5.3. Let $\mathcal{P} = \{S_1, \dots, S_K\}$ of $[T]$ be a fixed partition. For any globally shared representation Z_s and any collection of per-cluster representations $\{Z_{S_k}\}_{k=1}^K$, we apply CR identity once with $(Z, Y^{(1:T)}) = (Z_s, Y^{(1:T)})$ and once for each cluster with $(Z, Y^{(1:|S_k|)}) = (Z_{S_k}, Y^{(S_k)})$,

$$\sum_{t=1}^T I(Z_s; Y^{(t)}) = I(Z_s; Y^{(1:T)}) + \text{TC}(Y^{(1:T)}) - \text{TC}(Y^{(1:T)} | Z_s), \quad (\text{S4.33})$$

$$\sum_{t \in S_k} I(Z_{S_k}; Y^{(t)}) = I(Z_{S_k}; Y^{(S_k)}) + \text{TC}(Y^{(S_k)}) - \text{TC}(Y^{(S_k)} | Z_{S_k}). \quad (\text{S4.34})$$

Summing (S4.34) over $k = 1, \dots, K$ gives,

$$\sum_{t=1}^T I(Z_{S_{k(t)}}; Y^{(t)}) = \sum_{k=1}^K I(Z_{S_k}; Y^{(S_k)}) + \sum_{k=1}^K \text{TC}(Y^{(S_k)}) - \sum_{k=1}^K \text{TC}(Y^{(S_k)} | Z_{S_k}). \quad (\text{S4.35})$$

Subtracting (S4.33) from (S4.35) yields the *general gain identity*

$$\begin{aligned} \sum_{t=1}^T I(Z_{S_{k(t)}}; Y^{(t)}) - \sum_{t=1}^T I(Z_s; Y^{(t)}) &= \underbrace{\left(\text{TC}(Y^{(1:T)} | Z_s) - \sum_{k=1}^K \text{TC}(Y^{(S_k)} | Z_{S_k}) \right)}_{\text{difference in residual couplings}} \\ &\quad + \underbrace{\left(\sum_{k=1}^K \text{TC}(Y^{(S_k)}) - \text{TC}(Y^{(1:T)}) \right)}_{\text{redundancy term}} \\ &\quad + \underbrace{\left(\sum_{k=1}^K I(Z_{S_k}; Y^{(S_k)}) - I(Z_s; Y^{(1:T)}) \right)}_{\text{joint-predictive term}}. \end{aligned} \quad (\text{S4.36})$$

Now note that by Definition 5.2,

$$\sum_{k=1}^K \text{TC}(Y^{(S_k)}) - \text{TC}(Y^{(1:T)}) = -\text{TC}_{\text{between}}(\mathcal{P}).$$

Also define $\Delta_{\text{sh}} := \text{TC}(Y^{(1:T)} | Z_s)$ and $\Delta_k := \text{TC}(Y^{(S_k)} | Z_{S_k})$. Plugging these into (S4.36) gives

$$\sum_{t=1}^T I(Z_{S_{k(t)}}; Y^{(t)}) - \sum_{t=1}^T I(Z_s; Y^{(t)}) = \left(\Delta_{\text{sh}} - \sum_{k=1}^K \Delta_k \right) - \text{TC}_{\text{between}}(\mathcal{P}) + \left(\sum_{k=1}^K I(Z_{S_k}; Y^{(S_k)}) - I(Z_s; Y^{(1:T)}) \right). \quad (\text{S4.37})$$

Theorem 5.3 is stated for Z_s^* and $\{Z_{S_k}^*\}$ that are capacity optimal under budgets $\sum_k C_k = C$. In the capacity constrained case the tightness of bound is when the joint-predictive term saturates its information budget,

$$I(Z_s^*; Y^{(1:T)}) = C \quad \text{and} \quad I(Z_{S_k}^*; Y^{(S_k)}) = C_k \quad \forall k, \quad (\text{S4.38})$$

which holds in the Gaussian linear channel model when the encoder family can represent the optimal channel for each budget, and more generally whenever the optimum lies on the active constraint boundary. Under (S4.38),

$$\sum_{k=1}^K I(Z_{S_k}^*; Y^{(S_k)}) - I(Z_s^*; Y^{(1:T)}) = \sum_{k=1}^K C_k - C = 0.$$

Therefore, (S4.37) reduces exactly to

$$\sum_{t=1}^T I(Z_{S_{k(t)}}^*; Y^{(t)}) - \sum_{t=1}^T I(Z_s^*; Y^{(t)}) = \left(\Delta_{\text{sh}} - \sum_{k=1}^K \Delta_k \right) - \text{TC}_{\text{between}}(\mathcal{P}),$$

□

The clustered setting is strictly better than global sharing iff the right-hand side is strictly positive $\Delta_{\text{sh}} - \sum_{k=1}^K \Delta_k > \text{TC}_{\text{between}}(\mathcal{P})$. Without the saturation condition the exact gain identity consists of an additional joint-predictive difference term. In our empirical setup fixed-rank LoRA budget is typically budget active so that Equation (S4.38) is a good approximation, and the measured effect is dominated by the residual coupling and redundancy terms.

S5 PROOF OF THEOREM 4.4

Proof of Theorem 4.4. Using the squared-loss for a shared linear probe $\theta \in \mathbb{R}^d$, $\mathcal{L}_t(\theta) = \frac{1}{2} \mathbb{E}[(\theta^\top X - Y^{(t)})^2]$. Taking the derivative w.r.t θ ,

$$\nabla_{\theta} \mathcal{L}_t(\theta) = \mathbb{E}[(\theta^\top X - Y^{(t)})X] = \Sigma\theta - \mathbb{E}[Y^{(t)}X].$$

Under the model $Y^{(t)} = w_t^\top X + \varepsilon_t$ with $\varepsilon_t \perp X$ and $\mathbb{E}[\varepsilon_t] = 0$,

$$\mathbb{E}[Y^{(t)}X] = \mathbb{E}[(w_t^\top X)X] = \Sigma w_t.$$

Hence, at initialization $\theta_0 = 0$,

$$g_t = \nabla_{\theta} \mathcal{L}_t(\theta_0) = -\Sigma w_t.$$

Therefore the raw population gradient cosine similarity is

$$G_{ts}^{\text{raw}} = \cos(g_t, g_s) = \frac{g_t^\top g_s}{\|g_t\| \|g_s\|} = \frac{w_t^\top \Sigma^2 w_s}{\|\Sigma w_t\| \|\Sigma w_s\|},$$

which establishes the raw-gradient expression.

Let $\tilde{X} = \Sigma^{-1/2} X$, so $\tilde{X} \sim \mathcal{N}(0, I)$, and let us define $\tilde{w}_t = \Sigma^{1/2} w_t$. Then $Y^{(t)} = \tilde{w}_t^\top \tilde{X} + \varepsilon_t$ and

$$\text{Cov}(Y^{(t)}, Y^{(s)}) = \tilde{w}_t^\top \tilde{w}_s.$$

Moreover, the natural gradient, equivalently the gradient at $\theta_0 = 0$ in the whitened coordinates, equals $-\tilde{w}_t$, so the whitened-gradient cosine is

$$G_{ts}^{\text{nat}} = \frac{\tilde{w}_t^\top \tilde{w}_s}{\|\tilde{w}_t\| \|\tilde{w}_s\|}.$$

Since the denominators in the label correlation

$$\rho_{ts} = \frac{\text{Cov}(Y^{(t)}, Y^{(s)})}{\sqrt{\text{Var}(Y^{(t)}) \text{Var}(Y^{(s)})}}$$

are strictly positive, $\text{sign}(\rho_{ts}) = \text{sign}(\text{Cov}(Y^{(t)}, Y^{(s)})) = \text{sign}(\tilde{w}_t^\top \tilde{w}_s) = \text{sign}(G_{ts}^{\text{nat}})$.

Ordering under matched noise and normalized signal. Assume matched noise $\sigma_t^2 = \sigma^2$ for all t and equal signal power $\tilde{w}_t^\top \tilde{w}_t = \alpha^2$ for all t (equivalently $w_t^\top \Sigma w_t$ constant across tasks). Then

$$\text{Var}(Y^{(t)}) = \alpha^2 + \sigma^2 \quad \text{for all } t,$$

so the correlation simplifies to

$$\rho_{ts} = \frac{\tilde{w}_t^\top \tilde{w}_s}{\alpha^2 + \sigma^2}.$$

Thus, for any two pairs (t, s) and (t', s') ,

$$|\rho_{ts}| > |\rho_{t's'}| \iff |\tilde{w}_t^\top \tilde{w}_s| > |\tilde{w}_{t'}^\top \tilde{w}_{s'}|.$$

Under the same equal-norm condition, the cosine similarity satisfies

$$|G_{ts}^{\text{nat}}| = \frac{|\tilde{w}_t^\top \tilde{w}_s|}{\|\tilde{w}_t\| \|\tilde{w}_s\|} = \frac{|\tilde{w}_t^\top \tilde{w}_s|}{\alpha^2}.$$

Hence $|\rho_{ts}|$ and $|G_{ts}^{\text{nat}}|$ induce the same ordering over pairs, proving (ii). (If task signal powers differ, G_{ts}^{nat} still tracks the numerator but correlations are additionally rescaled by the marginal variances; the equal-power assumption isolates the redundancy structure.)

Total correlation of Gaussian labels and clustering equivalence. Let $Y = (Y^{(1)}, \dots, Y^{(T)})$ be jointly Gaussian with covariance Σ_Y and diagonal marginal variances $D = \text{diag}(\sqrt{\Sigma_{Y,11}}, \dots, \sqrt{\Sigma_{Y,TT}})$. Then $\Sigma_Y = DRD$, where R is the correlation matrix. For a multivariate Gaussian,

$$\text{TC}(Y) = \sum_{t=1}^T H(Y^{(t)}) - H(Y) = \frac{1}{2} \log \frac{\prod_{t=1}^T \Sigma_{Y,tt}}{\det(\Sigma_Y)}.$$

Using $\det(\Sigma_Y) = \det(D)^2 \det(R) = (\prod_t \Sigma_{Y,tt}) \det(R)$ yields

$$\text{TC}(Y) = -\frac{1}{2} \log \det(R).$$

Thus R fully determines $\text{TC}(Y)$. Under conditions in (ii) G_{ts}^{nat} is a strictly monotone rescaling of ρ_{ts} (and hence of R_{ts}) on $[0, 1]$. Similarly for absolute similarities when using $|\cdot|$. Therefore any clustering rule whose decisions depend only on the ordering of pairwise similarities produces the same partition when applied to G^{nat} as when applied to R . The raw-gradient cosine G^{raw} coincides with this quantity after whitening or when Σ is isotropic; otherwise the difference is governed by the anisotropy of Σ . This proves (iii). $\square$

S6 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

S6.1 DATASETS AND EVALUATION METRICS

GoEmotions. We select the twelve most frequent labels in the training split and treat each as a binary task. For evaluation we compute AUROC per label on the validation split and report the macro-average across all tasks. We also report per-label AUC in Section S7.2.

GLUE8. We evaluate on SST-2, MRPC, RTE, QNLI, QQP, MNLI (matched), CoLA, and STS-B. We report standard metrics, accuracy for SST-2/RTE/QNLI/MNLI, $(\text{Acc} + \text{F1})/2$ for MRPC/QQP and Matthews correlation coefficient (MCC) for CoLA, and $(\rho_{\text{Pearson}} + \rho_{\text{Spearman}})/2$ for STS-B. We report the average of task scores as the GLUE8 score.

S6.2 MODELS AND PARAMETER-EFFICIENT FINE-TUNING

All experiments use `bert-base-uncased` as the frozen encoder. We attach LoRA adapters to the attention query/value projections and only LoRA parameters and task heads are trainable. For LoRA we use $\alpha = 16$ with a dropout probability of 0.05.

We control parameterization through a total rank budget r_{tot} . In the shared setting, the single adapter has rank r_{tot} . In clustered routing with K adapters, we allocate ranks across adapters such that $\sum_{k=1}^K r_k = r_{\text{tot}}$ with $r_k \geq 1$. For GLUE8 we use proportional rank allocation (larger clusters receive more rank), while GoEmotions uses fixed-sum allocation consistent with the split runner.

Model training. We train with AdamW, FP16 enabled, linear warmup (6% of total steps), and gradient clipping (norm 1.0). We use batch size of 16 and steps-per-epoch 1400 for 3 epochs. We use weight decay on task heads but set it to 0 for LoRA weights. This is done to avoid decoupled weight decay shrinking inactive adapters in multi-adapter setting.

S6.3 SIMILARITY ESTIMATION AND CLUSTERING

We first partition tasks into K clusters and use one adapter per cluster with ranks $\{r_k\}$ such that $\sum_k r_k = R$ and $r_k \geq 1$. We estimate task similarity without training full adapters by attaching a probe LoRA adapter of rank 1, denoted a_{probe} , and computing per-task gradient vectors,

$$g_t := \frac{1}{B} \sum_{b=1}^B \nabla_{\theta_{\text{probe}}} \ell_t(\theta_{\text{probe}}; \text{batch } b), \quad \tilde{g}_t := \frac{g_t}{\|g_t\|_2}.$$

where for similarity we use cosine $s_{ij} = \langle \tilde{g}_i, \tilde{g}_j \rangle$. We use agglomerative clustering on the distance matrix $D_{ij} = 1 - s_{ij}$ with average linkage and fixed K .

For GoEmotions we use probe rank 1, 5 similarity seeds, and 20 gradient-accumulation steps per task, and for GLUE8 we use 30 steps per task. We warm-start task heads for 200 steps before computing gradients to stabilize the probe. Given the mean similarity matrix, we compute an average-linkage agglomerative clustering and select K using a gap criterion constrained to $K \in [2, 6]$.

S6.4 CLUSTERING-METHOD SENSITIVITY

To test sensitivity to the clustering design, we sweep the number of clusters $K \in \{2, 3, 4, 5, 6\}$ and compare average-linkage, complete-linkage, and spectral clustering at ranks 8 and 12. Complete linkage gives the best AUC at the strongest setting ($K = 6, r = 8$), while average linkage tracks it closely and obtains the lowest residual-coupling value at $K = 6, r = 12$. Across algorithms, $\hat{\Delta}$ generally decreases as K grows, consistent with the prediction that more specialized routes reduce residual conflict.

S6.5 ADDITIONAL RESULTS

S7 LINEAR-GAUSSIAN VERIFICATION OF THE CR INEQUALITY

We verify the CR inequality in a controlled linear-Gaussian setting where all quantities are computable in closed form. We sample $X \sim \mathcal{N}(0, I_d)$ ($d = 20$) and define T task labels

$$Y^{(t)} = a_t^\top X + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad (\text{S7.39})$$

with independent $\{\varepsilon_t\}_{t=1}^T$ and $\sigma_\varepsilon = 0.1$. Let $A \in \mathbb{R}^{T \times d}$ stack $a_t^\top$ as rows. By varying the alignment of $\{a_t\}$ we sweep label redundancy. If $a_i \approx a_j$ then labels share signal and $\text{TC}(Y^{1:T})$ increases, while near-orthogonality yields $\text{TC}(Y^{1:T}) \approx 0$.

Representations are produced by a rank- r linear encoder

$$Z = W^\top X + \eta, \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2 I_r), \quad (\text{S7.40})$$

Table S6.5: Clustering sensitivity on GoEmotions: mean $\pm$ std over five random seeds.

Algorithm	K	Rank	avg_auc	$\hat{\Delta}$
avg	2	8	0.6950 $\pm$ 0.0602	0.2331 $\pm$ 0.0286
avg	2	12	0.7011 $\pm$ 0.0729	0.2305 $\pm$ 0.0298
avg	3	8	0.7149 $\pm$ 0.0368	0.2229 $\pm$ 0.0145
avg	3	12	0.7021 $\pm$ 0.0559	0.2088 $\pm$ 0.0124
avg	4	8	0.8004 $\pm$ 0.0087	0.2341 $\pm$ 0.0089
avg	4	12	0.7817 $\pm$ 0.0174	0.2155 $\pm$ 0.0102
avg	5	8	0.8065 $\pm$ 0.0182	0.2290 $\pm$ 0.0093
avg	5	12	0.8081 $\pm$ 0.0228	0.2065 $\pm$ 0.0086
avg	6	8	0.8133 $\pm$ 0.0177	0.2074 $\pm$ 0.0091
avg	6	12	0.8101 $\pm$ 0.0208	0.2013 $\pm$ 0.0095
complete	2	8	0.7005 $\pm$ 0.0542	0.2340 $\pm$ 0.0271
complete	2	12	0.7092 $\pm$ 0.0575	0.2298 $\pm$ 0.0260
complete	3	8	0.8057 $\pm$ 0.0172	0.2437 $\pm$ 0.0102
complete	3	12	0.7945 $\pm$ 0.0180	0.2295 $\pm$ 0.0099
complete	4	8	0.7900 $\pm$ 0.0187	0.2322 $\pm$ 0.0103
complete	4	12	0.7882 $\pm$ 0.0141	0.2095 $\pm$ 0.0094
complete	5	8	0.8144 $\pm$ 0.0144	0.2272 $\pm$ 0.0095
complete	5	12	0.8067 $\pm$ 0.0208	0.2087 $\pm$ 0.0090
complete	6	8	0.8392 $\pm$ 0.0091	0.2159 $\pm$ 0.0083
complete	6	12	0.8355 $\pm$ 0.0117	0.2028 $\pm$ 0.0089
spectral	2	8	0.7598 $\pm$ 0.0296	0.2688 $\pm$ 0.0181
spectral	2	12	0.7632 $\pm$ 0.0271	0.2632 $\pm$ 0.0167
spectral	3	8	0.7799 $\pm$ 0.0186	0.2528 $\pm$ 0.0124
spectral	3	12	0.7546 $\pm$ 0.0474	0.2400 $\pm$ 0.0146
spectral	4	8	0.7824 $\pm$ 0.0139	0.2173 $\pm$ 0.0098
spectral	4	12	0.7899 $\pm$ 0.0076	0.2296 $\pm$ 0.0093
spectral	5	8	0.8124 $\pm$ 0.0119	0.2040 $\pm$ 0.0086
spectral	5	12	0.8057 $\pm$ 0.0094	0.2144 $\pm$ 0.0089
spectral	6	8	0.8033 $\pm$ 0.0083	0.2154 $\pm$ 0.0084
spectral	6	12	0.8008 $\pm$ 0.0127	0.2062 $\pm$ 0.0088

with $\sigma_\eta = 0.5$. If $\text{row}(W) \supseteq \text{row}(A)$, then AX is a linear function of Z , i.e. there exists M such that $AX = MZ$, and hence

$$Y = AX + \varepsilon = MZ + \varepsilon. \quad (\text{S7.41})$$

Because ε has independent coordinates, the tasks are conditionally independent given Z , implying $\text{TC}(Y^{1:T} \mid Z) = 0$, so the CR inequality becomes tight.

When the representation is rank-limited, the residual coupling term is exactly the conditional total correlation $\text{TC}(Y^{1:T} \mid Z)$, which quantifies the remaining dependence among tasks after conditioning on the representation. We compute $\text{TC}(Y^{1:T})$ and $\text{TC}(Y^{1:T} \mid Z)$ from Gaussian covariance identities (log-determinant forms), and compute $I(Z; Y^{(t)})$ via conditional-variance formulas. Figure S7.6 reports sweeps across redundancy level, comparing optimal vs random vs misaligned encoders, number of tasks at low redundancy to illustrate negative transfer in average per-task information, and rank r to show saturation once the encoder spans the task subspace.

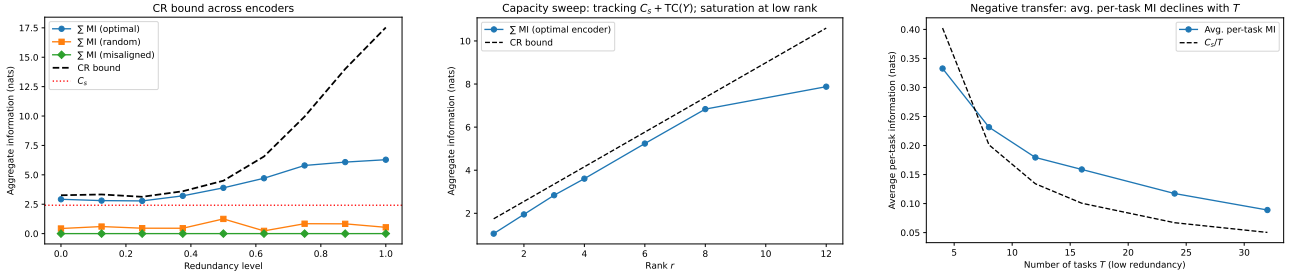


Figure S7.6: Left: CR holds across encoder types, with tightness for the optimal encoder and slack for random or misaligned ones. Middle: Capacity sweep shows aggregate information tracks $C_s + TC(Y)$, saturating at low rank. Right: Negative transfer demonstrated, as average per-task information declines when the number of tasks increases under fixed capacity and low redundancy.

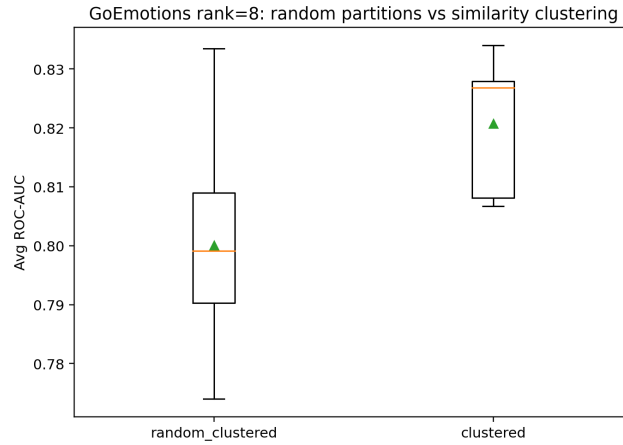


Figure S7.7: **GoEmotions (rank=8): clustered vs random partitions (available baseline)**. Distribution of macro-average ROC-AUC over 45 random partitions (seed 41) compared to similarity-based clustering. We report this as a sanity check rather than a universal significance claim because random baselines are only available at rank 8 in the split-full runs.

S7.1 GOEMOTIONS: CLUSTERED VS RANDOM PARTITIONS (AVAILABLE BASELINE)

S7.2 PER-EMOTION GAINS AT RANK 12

S7.3 GOEMOTIONS BASELINES

We include additional baselines from the earlier harness using PCGrad, GradNorm, single-task oracle. Here, the single-task baseline is an oracle upper bound and is not compute-matched.

S7.4 SIMILARITY HEATMAPS

We make a note that clustered/private settings require $r_{\text{tot}} \geq K$ (and private additionally requires $r_{\text{tot}} \geq T$). For GLUE8 and GoEmotions, $K = 6$ in our clustering, hence routing is only feasible for rank budgets ≥ 8 in the plotted ranges.

Table S7.6: Clustered vs random-clustered (empirical p-value and effect size).

rank	clustered_mean	random_mean	p_emp	cohen_d	n_clustered	n_random
8.0000	0.821	0.800	0.0889	1.499	5.0000	45.0000

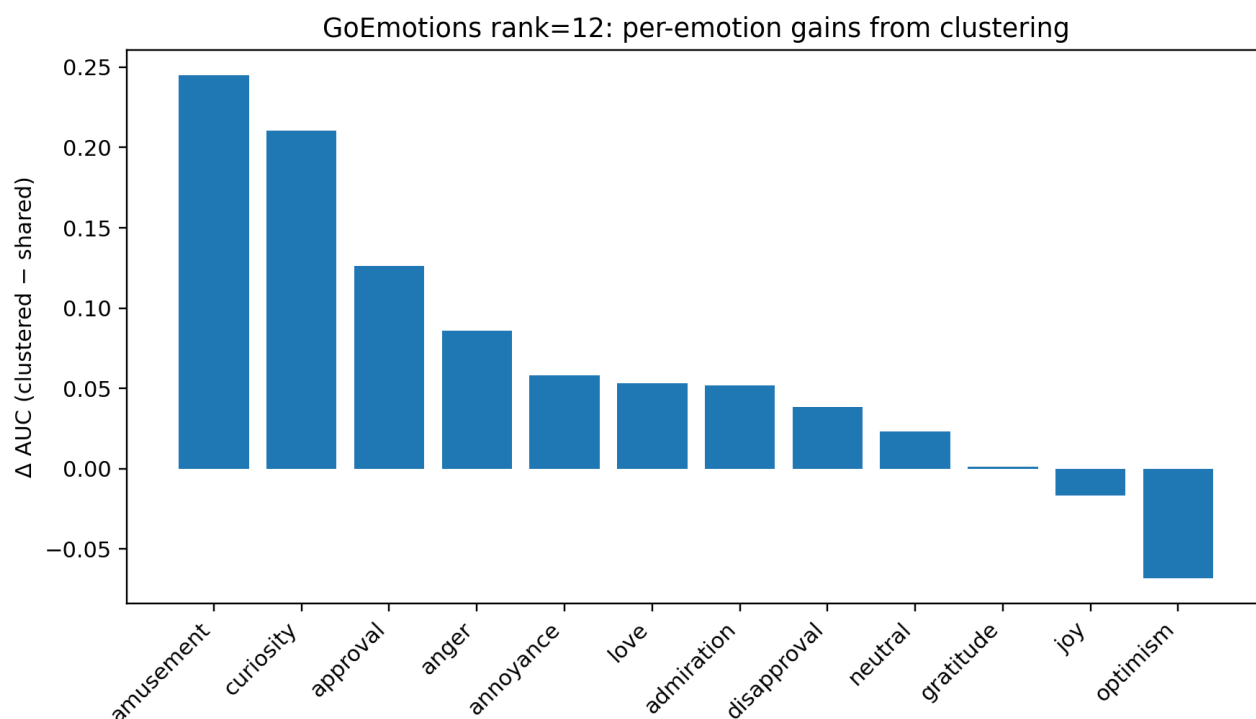


Figure S7.8: **GoEmotions per-label effect of clustering.** Difference in AUROC per emotion (clustered minus shared). Improvement in performance is on several harder labels (e.g., amusement, curiosity), while some labels change minimally. Error bars are reported in Table S7.7.

Table S7.7: Per-emotion AUROC for shared vs clustered (mean $\pm$ CI95) and their difference.

task	shared_mean	shared_ci95	clustered_mean	clustered_ci95	delta_mean
amusement	0.564	0.093	0.809	0.021	0.245
curiosity	0.675	0.094	0.886	0.017	0.211
approval	0.494	0.047	0.620	0.045	0.126
anger	0.753	0.074	0.839	0.033	0.086
annoyance	0.727	0.064	0.785	0.043	0.058
love	0.890	0.026	0.944	0.011	0.053
admiration	0.848	0.011	0.900	0.024	0.052
disapproval	0.704	0.047	0.742	0.040	0.038
neutral	0.761	0.012	0.784	0.009	0.023
gratitude	0.982	0.006	0.984	0.004	0.002
joy	0.800	0.008	0.784	0.055	-0.016
optimism	0.653	0.041	0.585	0.136	-0.068

Table S7.8: Additional baselines across available ranks (mean $\pm$ CI95).

rank	regime	n	avg_auc_mean	avg_auc_ci95
4	shared	5	0.725	0.032
4	clustered	0	—	—
4	private	0	—	—
4	single_task	0	—	—
4	pcgrad_shared	5	0.903	0.003
4	gradnorm_shared	5	0.894	0.001
4	random_clustered	0	—	—
8	shared	5	0.702	0.081
8	clustered	0	—	—
8	private	0	—	—
8	single_task	0	—	—
8	pcgrad_shared	5	0.904	0.003
8	gradnorm_shared	5	0.901	0.002
8	random_clustered	0	—	—
12	shared	5	0.738	0.012
12	clustered	5	0.827	0.005
12	private	5	0.830	0.003
12	single_task	5	0.929	0.000
12	pcgrad_shared	0	—	—
12	gradnorm_shared	0	—	—
12	random_clustered	55	0.824	0.003

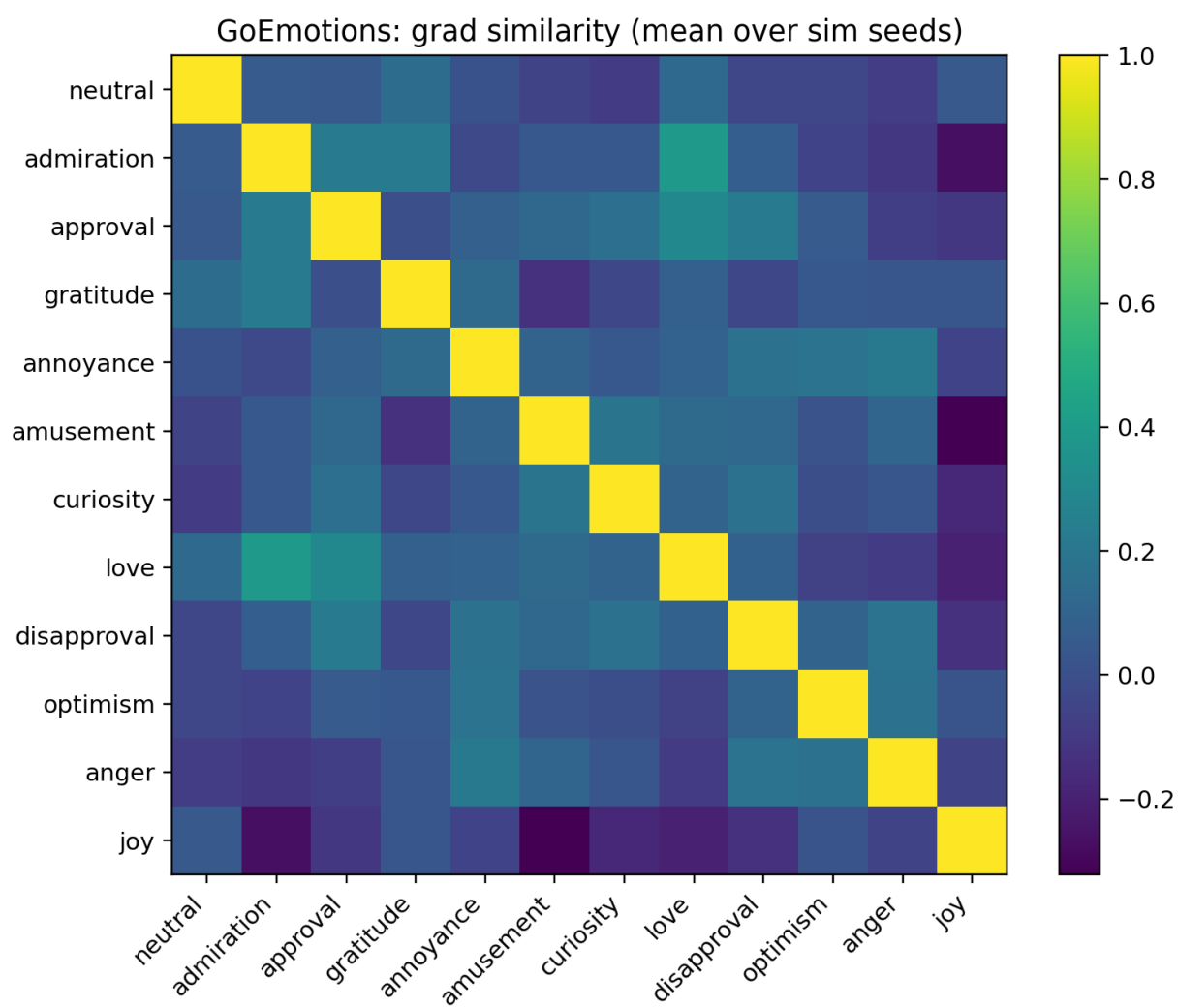


Figure S7.9: **Gradient similarity heatmap for GoEmotions.** Mean cosine similarity of probe gradients across five similarity seeds. The resulting clusters group semantically related labels.



Figure S7.10: **Gradient similarity heatmap for GLUE** Mean cosine similarity of probe gradients across 5 similarity seeds. The clustering isolates sentiment (SST-2) and acceptability (CoLA) while grouping entailment/paraphrase-style tasks (e.g., MRPC–RTE and QNLI–MNLI).