COBRA: Contribution-Based Bayesian Rank Allocation for Parameter-Efficient Fine-Tuning

Hongcheng Ding[*], **Xuanze Zhao**[*†], Liu Xuanhuang, Jing Jin, Shamsul Nahar Abdullah, Deshinta Arrova Dewi

Dongfang College, Zhejiang University of Finance & Economics · INTI International University · SEGi University

[*]Equal Contribution [†]Corresponding Author 20070032@zufedfc.edu.cn, I24025877@sudent.newinti.edu.my

▶ Abstract

Full fine-tuning of large language models (LLMs) incurs prohibitive computational and storage costs. Parameter-efficient fine-tuning (PEFT) addresses this limitation, with Low-Rank Adaptation (LoRA) gaining widespread adoption due to its simplicity and zero inference overhead. However, LoRA and its variants typically rely on uniform rank allocation or a single importance metric such as gradient magnitude or output sensitivity to guide rank distribution. This approach fails to recognize that gradient magnitude and output contribution are decoupled properties, leading to suboptimal allocation where critical layers are under-provisioned while less important ones waste capacity.

We introduce **COBRA** (Conductance-based Bayesian Rank Allocation), it operates in three core stages to a principled framework integrating dual importance factors for adaptive rank allocation in PEFT.

S1 Layer Conductance: Path-integral attribution for layer contribution $\Phi^{(l)}$
S2 Dual-Factor Aggregation: Multiplicative fusion with gradient $\Gamma^{(l)} \rightarrow$ TA-LC prior $\pi^{(l)}$
S3 Bayesian Allocation: ELBO-driven variational rank optimization $\{r_l\}$

On GLUE, MMLU, GSM8K, HumanEval, and financial regression benchmarks, **COBRA** achieves **+1.6 pts** on GLUE, **+0.7%** over AdaLoRA, **+2.5%** over GoRA on LLaMA-7B, and **+6.6%** in high-rank regimes, under *identical* parameter budgets and with **zero inference overhead**.

P1 Uniform Rank Allocation: Static equal ranks neglect heterogeneous layer functional importance.
P2 Single-Metric Deficiency: Isolated metrics fail to capture decoupled layer contribution and adaptation demand.
P3 Parameter Capacity Mismatch: Heuristic rank allocation causes critical layer under-provisioning and parameter waste.

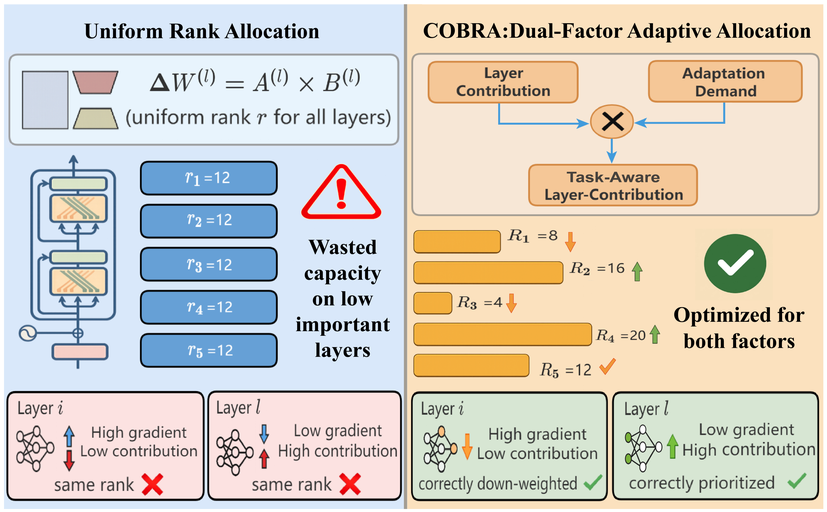

Figure 1: (Left) Uniform rank wastes capacity on low-contribution layers. **(Right)** COBRA's dual-factor adaptive allocation correctly prioritizes layers that are both *contributive* and *adaptation-demanding*.

▶ Motivation: Two Critical Gaps

We probe **7 architectures** (encoder/decoder/enc–dec, 86M–7B params) and uncover two failure modes in single-factor rank allocation:

Challenge #1: Layer Heterogeneity
- Gradient magnitude varies **9**×–**18**× across layers
- Output contribution varies **10**×–**19**×
- **Peaks diverge:** contribution peaks at *middle* layers (L13–16); gradient peaks at *top* layers (L21–24)
- ⇒ Uniform allocation systematically wastes budget

Challenge #2: Decoupled Characteristics
- Gradient ⊥ contribution: only *moderately* correlated (ρ=0.66 RoBERTa-large, ρ=0.56 LLaMA-2-7B)
- **Failure A:** high gradient + low contribution → rank wasted on minimal-impact layers (GoRA's blind spot)
- **Failure B:** high contribution + moderate gradient → critical layers under-provisioned (AdaLoRA's blind spot)
- ⇒ *Neither metric alone suffices*

These gaps motivate a **dual-factor, Bayesian** framework that integrates contribution and adaptation demand into a single principled allocation rule.

▶ Key Contributions

C1. Decoupling Diagnosis: First systematic evidence that gradient magnitude and output contribution are decoupled across 7 transformer architectures, exposing the root cause of single-factor allocation failure.

C2. Interpretability-Guided PEFT: A novel *Task-Adaptive Layer Conductance* (TA-LC) distribution combining path-integral attribution with adaptation demand, embedded as Bayesian prior in a variational rank-allocation objective.

C3. Strong, Scalable Gains: **+1.6** GLUE (RoBERTa-base), **+0.7%** over AdaLoRA, **+5.9%** over LoRA on LLaMA-7B, **+6.6%** in high-rank regimes; advantage *grows* with parameter budget.

▶ Method: COBRA Framework

COBRA jointly leverages **two complementary signals** unified via Bayesian inference:

$\Phi^{(l)}$: Layer Contribution → $\Gamma^{(l)}$: Adaptation Demand → $\pi^{(l)}$: TA-LC Prior

Stage 1: Layer Conductance Attribution Quantify exhaustive, non-redundant layer contribution via path-integral attribution:

$$\mathcal{A}^{(l)}(x) = \sum_{j=1}^{d_l} (h_j^{(l)}(x) - h_j^{(l)}(x_0)) \cdot \int_0^1 \frac{\partial F(x_\alpha)}{\partial h_j^{(l)}} d\alpha, \quad \Phi^{(l)} = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{x \in \mathcal{D}_{\text{cal}}} |\mathcal{A}^{(l)}(x)| \tag{1}$$

- **Calibration set:** $x_\alpha = x_0 + \alpha(x - x_0)$, 50 steps, $|\mathcal{D}_{\text{cal}}|$=1000
- **Completeness:** $\sum_l \mathcal{A}^{(l)}(x) = F(x) - F(x_0) \Rightarrow$ exhaustive & non-redundant
- One-time computation, amortized across all training runs

Stage 2: Dual-Factor Aggregation → TA-LC Fuse layer contribution and adaptation demand multiplicatively:

$$\Gamma^{(l)} = \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{cal}}}\left[\|\nabla_{W_0^{(l)}} \mathcal{L}(x,y)\|_F^2\right]}, \quad \tilde{\Phi}^{(l)} = \frac{\Phi^{(l)}}{\sum_{l'=1}^{L} \Phi^{(l')}}, \quad \tilde{\Gamma}^{(l)} = \frac{\Gamma^{(l)}}{\sum_{l'=1}^{L} \Gamma^{(l')}}, \quad \Psi^{(l)} = \tilde{\Phi}^{(l)} \cdot \tilde{\Gamma}^{(l)} \quad \pi^{(l)} = \frac{\Psi^{(l)}}{\sum_{l'=1}^{L} \Psi^{(l')}}, \tag{2}$$

- **Multiplicative** (not additive): non-compensatory -> both factors must matter
- Down-weights layers that score high on *only one* factor (Failure A & B)
- Yields a normalized probability distribution over layers with $\sum_{l=1}^{L} \pi^{(l)} = 1$.

Stage 3: Bayesian Rank Allocation

$$\Delta W^{(l)} = \sum_{k=1}^{r_{\max}} z_k^{(l)} \mathbf{a}_k^{(l)} (\mathbf{b}_k^{(l)})^\top, \qquad p(z_k^{(l)}=1) = \rho^{(l)} = \sigma\big(\gamma\,(\pi^{(l)} - \bar{\pi})\big) \tag{3}$$

- Rank-level gates $z_k^{(l)} \in \{0,1\}$; TA-LC-informed Bernoulli prior
- Continuous relaxation + Gumbel reparameterization for end-to-end gradients
- Top-R_{total} thresholding extracts deterministic $\{r_l\}$ under budget

Variational Objective (ELBO):

$$\mathcal{J}_{\text{ELBO}} = \underbrace{\mathbb{E}_q[\log p(\mathcal{D} \mid \theta, \tilde{\mathbf{z}})]}_{\text{task likelihood}} - \beta \cdot \underbrace{D_{\text{KL}}(q_\phi(\tilde{\mathbf{z}}) \,\|\, p(\mathbf{z}))}_{\text{TA-LC regularization}} \tag{4}$$

$\theta = \{A^{(l)}, B^{(l)}\}$: LoRA params; $\phi = \{\alpha_k^{(l)}\}$: variational params; β=0.01, γ=5.0, λ: 1.0→0.1 (annealed).

Key Design Advantages:
- **Dual-factor:** fixes gradient–contribution decoupling
- **Bayesian:** smooth optimization, no hard pruning instability
- **Zero inference overhead:** same as standard LoRA
- **Amortized calibration:** one-time cost reused across runs

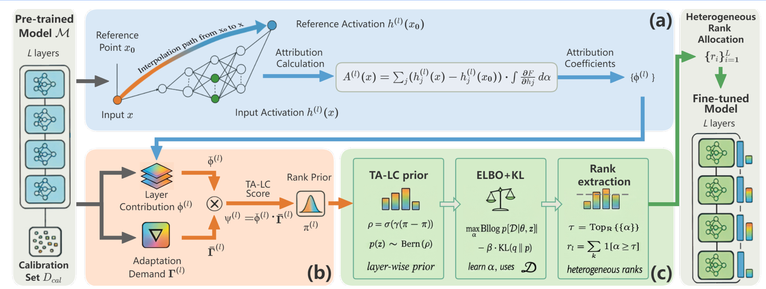

Figure 2: (a) Conductance attribution via interpolation path. **(b)** Multiplicative fusion of contribution & adaptation demand into TA-LC prior. **(c)** ELBO-based Bayesian gating with threshold extraction → heterogeneous ranks $\{r_l\}_{l=1}^{L}$.

▶ Training Protocol & Implementation

Algorithm 1 COBRA Procedure

Require: Pre-trained model $\mathcal{M}$, calibration set $\mathcal{D}_{\text{cal}}$, training set $\mathcal{D}$, budget R_{total}, max rank $r_{\max}$
// **Stage 1: Dual-Factor Estimation**
Compute $\{\Phi^{(l)}\}$ and $\{\Gamma^{(l)}\}$ via Eq. (1)–(2)
// **Stage 2: Prior Construction**
Compute $\{\pi^{(l)}\}$ and $\{\rho^{(l)}\}$ via Eq. (2)–(3)
// **Stage 3: Bayesian Training**
Initialize $A^{(l)} \sim \mathcal{N}$, $B^{(l)} \leftarrow 0$, $\alpha_k^{(l)} \leftarrow \rho^{(l)}$
for each training iteration **do**
  Sample $\tilde{z}_k^{(l)}$ via Eq. (10)
  Compute $\mathcal{J}_{\text{ELBO}}$ via Eq. (4)
  Update θ, ϕ via gradient descent
end for
// **Stage 4: Rank Extraction**
$\tau \leftarrow \text{Top}_{R_{\text{total}}}(\{\alpha_k^{(l)}\}_{l,k})$
for $l = 1$ to L **do**
  $r_l \leftarrow \sum_{k=1}^{r_{\max}} \mathbb{1}[\alpha_k^{(l)} \geq \tau]$
end for
Ensure: Fine-tuned model with $\{r_l\}_{l=1}^{L}$

Hyperparameters & Setup:
- Adapters on *all* linear layers (attention + FFN)
- AdamW, lr $\in \{1, 2, 5\} \times 10^{-4}$, batch 16/32, 6% warmup, linear decay
- Rank budgets: $R_{\text{total}} \in \{192, 384, 768\}$ (NLU) & 9K, 12K (finance)
- Hardware: NVIDIA A100 80GB, FP16 activation / FP32 gradient

▶ Evidence: Gradient ≠ Contribution

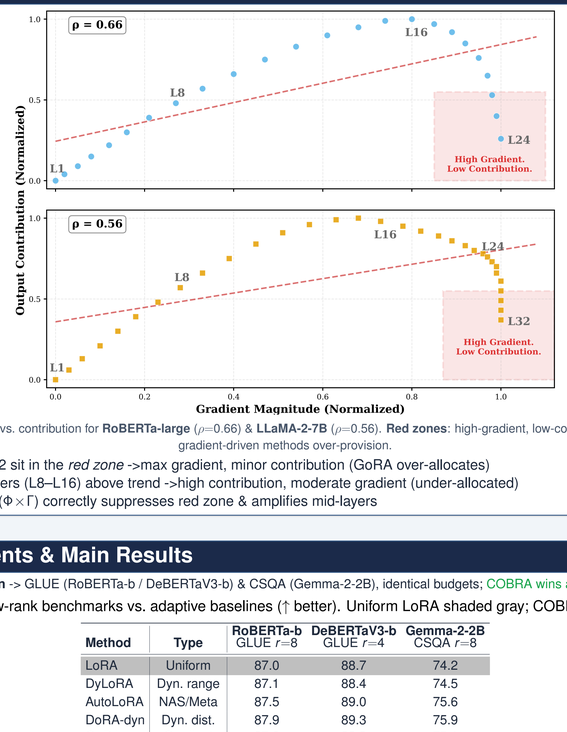

Figure 3: Gradient vs. contribution for **RoBERTa-large** (ρ=0.66) & **LLaMA-2-7B** (ρ=0.56). **Red zones**: high-gradient, low-contribution layers that gradient-driven methods over-provision.

- **Failure A:** L24/L32 sit in the *red zone* ->max gradient, minor contribution (GoRA over-allocates)
- **Failure B:** Mid-layers (L8–L16) above trend ->high contribution, moderate gradient (under-allocated)
- **COBRA's TA-LC** ($\Phi \times \Gamma$) correctly suppresses red zone & amplifies mid-layers

▶ Experiments & Main Results

NLU & generalization -> GLUE (RoBERTa-b / DeBERTaV3-b) & CSQA (Gemma-2-2B), identical budgets; COBRA wins all three:

Table: Low-rank benchmarks vs. adaptive baselines (↑ better). Uniform LoRA shaded gray; COBRA in pink.

Method	Type	RoBERTa-b GLUE r=8	DeBERTaV3-b GLUE r=4	Gemma-2-2B CSQA r=8
LoRA	Uniform	87.0	88.7	74.2
DyLoRA	Dyn. range	87.1	88.4	74.5
AutoLoRA	NAS/Meta	87.5	89.0	75.6
DoRA-dyn	Dyn. dist.	87.9	89.3	75.9
GoRA	Gradient	87.6	88.8	75.1
COBRA	**Dual-factor**	**88.6**	**89.9**	**76.1**

LLM scaling (vs. gradient-driven GoRA, LLaMA-7B, r_0=8):
- MMLU **48.6** (+1.5), GSM8K **45.9** (+2.1), HumanEval **52.7** (+3.7)
- Relative gain over LoRA **+5.9%** vs. GoRA's +3.4%

High-rank regression financial sentiment, 3 datasets:

Table: MSE↓ across rank budgets. LoRA gray (baseline); COBRA pink. Gains grow with budget (avg. **+6.6%**).

Method	Config.	NEU MSE	NEU MAE	FXE MSE	FXE MAE	INV MSE	INV MAE
Full FT	-	0.0189	0.0981	0.0202	0.1014	0.0183	0.0959
LoRA	r=384	0.0189	0.0977	0.0205	0.1021	0.0184	0.0961
	r=512	0.0186	0.0958	0.0199	0.1007	0.0183	0.0955
AdaLoRA	640→512	0.0218	0.1052	0.0231	0.1079	0.0196	0.1005
COBRA	r_{avg}=384	0.0177	0.0940	0.0192	0.0978	0.0172	0.0917
	r_{avg}=512	**0.0174**	**0.0931**	**0.0185**	**0.0958**	**0.0170**	**0.0908**

▶ Ablation Study (DeBERTaV3-base)

Factor Aggregation Variant	Acc.	Prior γ (inverted-U) Variant	Acc.	KL Weight β Variant	Acc.
Φ/Γ only	88.8	γ=0 (none)	88.5	β=0	88.8
Φ+Γ (add.)	89.2	γ=50 (over)	88.3	β=1.0 (over)	88.2
$\Phi \times \Gamma$ (ours)	**89.9**	**γ=5.0 (ours)**	**89.9**	**β=0.01 (ours)**	**89.9**

- **Multiplicative** fusion is non-compensatory: a layer must score high on *both* factors (+0.7 over additive).
- Interpretability-guided prior is essential: weak → unguided, strong → over-constrained.

▶ Conclusion & Takeaways

COBRA resolves the gradient–contribution decoupling that limits prior PEFT by unifying **interpretability** (path-integral conductance) with **adaptation demand** under a **Bayesian** objective.

- **Dual-factor** allocation fixes single-metric blind spots (GoRA & AdaLoRA).
- **Bayesian relaxation** gives stable optimization, no hard-pruning instability.
- Gains **scale with budget** (+1.6 GLUE → +6.6% high-rank); **zero inference overhead**.
- **Model-agnostic**: encoders, decoders & encoder-decoders.

Future: Vision-language PEFT · Causal attribution · Auto-budget allocation.