arXiv:2601.00231v1 [cs.LG] 1 Jan 2026

- 采用 K-FAC 预条件、费舍尔引导的重新投影与动态秩自适应的几何感知参数高效微调

Pritish Saha[1] **Chandrav Rajbangshi**[1] **RudraGoyal**[1] **MohitGoyal**[1]
Anurag Deo[1] **Biswajit Roy**[1] **Ningthoujam Dhanachandra Singh**[1]
RaxitGoswami[1] **Amitava Das**[2]
[1]RAAPID Lab，美国 [2]Pragya Lab，BITS Pilani，果阿

摘要

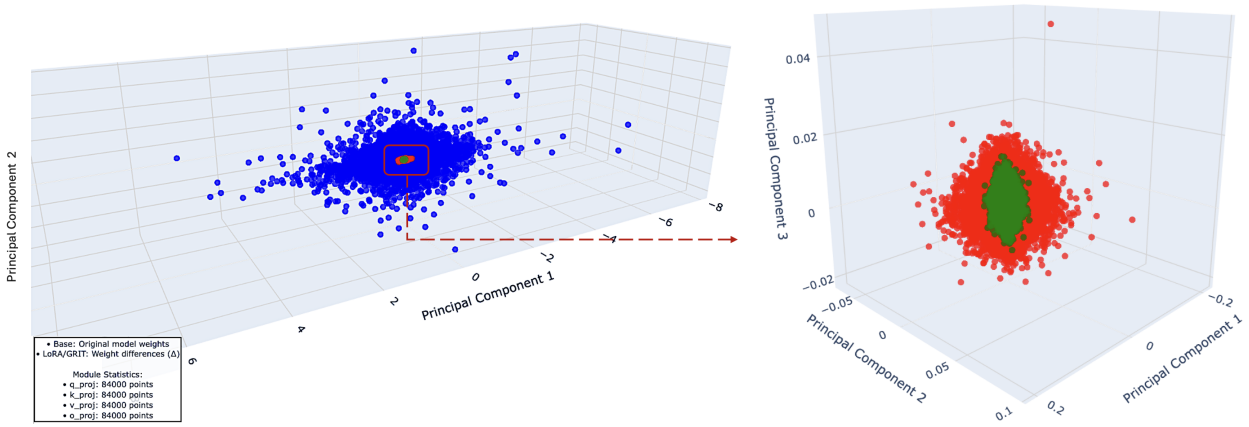
参数高效微调（PEFT）已成为将大语言模型适配到特定领域和应用场景的默认途径，但 LoRA 和 QLoRA 等广泛使用的方法在很大程度上仍未考虑几何结构：它们使用一阶下降，在固定且随机定向的低秩子空间中进行优化，基本忽略了局部损失曲率。这会扩大有效更新预算，并加剧沿约束较弱方向的漂移。我们提出 GRIT，一种动态的曲率感知 LoRA 方法。GRIT 保留 LoRA 参数化形式，同时：（1）使用 K-FAC 作为自然梯度代理，在秩空间中对梯度进行预条件；（2）定期将低秩基重新投影到主导 Fisher 特征方向上，以抑制漂移；（3）根据谱调整有效秩，使容量集中在信号所在之处。最终效果是，在使用更少有效参数的同时，引导更新朝高信号、低干扰的方向移动。在 LLaMA 骨干模型上的指令遵循、理解和推理基准测试中，GRIT 的表现与 LoRA/QLoRA 相当或更优，同时平均减少 ∼46%的可训练参数（不同任务间为 25–80%），且在多种提示风格和不同数据组合下，实际质量并未下降。由于微调常会导致灾难性遗忘，我们用曲率调制幂律 $_L$GRIT $_{pt} = L^0_{pt} + A \frac{D^\beta_{ft}}{(\Xi\text{GRIT}_N)^\alpha} + E$ with $\Xi_{\text{GRIT}} = (1 + \gamma_r r_{\text{eff}})(1 + \gamma_a \rho_{\text{align}})(1 + \gamma_p \pi_{\text{proj}})$ 对 GRIT 的漂移进行建模，以刻画有效秩、Fisher 对齐和投影保真度；实证结果表明，GRIT 的漂移始终低于 LoRA，并在参数更新与保留率的权衡前沿上优于强大的 PEFT 方法和优化器基线（例如 Orthogonal-LoRA、IA^3, DoRA/高效微调、Shampoo）。代码仓库。

GRIT — 一览

W 简而言之：**GRIT 是一种几何感知*LoRA*方法，沿着测得的曲率集中进行低秩适配**，并通过**K-FAC秩空间预条件化、费舍尔引导的重新投影**和**动态秩**，在保持质量的同时减少更新占用。⚙ 核心机制：GRIT 会将 LoRA 因子**周期性重新投影**到**Fisher特征值最高的特征方向**上，并使用**累积特征质量规则**（配合界限和滞后）动态选择**有效秩**。j 更新占用：在图示的深度策略中（冻结第 1–10 层，适配第 11–32 层），与 LoRA 和 LoRA+K-FAC 相比，GRIT 降低了**平均更新密度**和更新参数总数，从而使各层的适配占用更加紧凑。🕐 实际开销：GRIT 将重型操作保留在 $r \times r$中，并且只触发**少量重新投影**（在计时设置中，每 1k 步约**1.8–2.4 次**），平均每步耗时开销**仅为个位数百分比**，且**P99 峰值**与重投影事件相对应。¤ 记录/审计哪些内容：附录以**有效秩** r_{eff}、**对齐重叠度**ρ_{align}和**保留质量**π_{proj}为变量，对**有效容量乘数** Ξ_{GRIT} 进行了形式化定义；这些可操作的遥测指标可用于验证“几何感知压缩”是否确实发生。 . 局限性（务必保持审慎）：性能取决于**曲率估计质量**（K-FAC 假设、早期 Fisher噪声），也取决于**投影频率**旋钮 T_{proj}；调整该旋钮需要在稳定性与计算量之间权衡。

1 参数高效微调的盲点：学习惯性与灾难性遗忘

适配十亿参数的大语言模型会给内存和带宽带来压力；参数高效微调通过冻结大部分权重、仅训练一小部分参数（例如低秩更新）来应对这一问题。在参数高效微调的各种变体中，**LoRA** [Hu 等人、 2021] 以及**QLoRA** [Dettmers 等人、 2023] 已成为事实上的标准。**新近显现的权衡**：近期证据表明，“***LoRA*** 学得较少，遗忘也较少”[Biderman 等人， 2024]这说明，相较于全量微调，LoRA 的低秩更新预算往往会带来较小的困难目标增益（例如代码/数学），同时保留更多基础模型的广泛能力（例如常识和通用语言能力）。在



(a) 全局视图：原始模型权重与 LoRA/GRIT Δ-权重**(b)** 放大原点附近的视图，在 PCA 空间中比较 LoRA 与 GRIT Δ-权重。

图 1：参数更新的几何结构：GRIT 将 Δ 权重集中到曲率对齐子空间中。设置。对跨层注意力投影（q_{proj}、k_{proj}、v_{proj}、o_{proj}）的参数向量进行主成分分析；点经均值中心化后嵌入到主成分中（不做额外缩放）。(a) 整体（左）。蓝点表示由原始（基础）权重构成的整个模型参数空间。叠加在中心处的红点表示 LoRA 更新增量（ΔW），绿点表示 GRIT 更新增量。图中显示，完整的基础空间范围广且具有各向异性，而 LoRA 和 GRIT 都在小得多的中心区域——即有效的微调流形——内运行。(b) 放大视图 Δ（右）。省略基础权重点云，以便直接比较更新：红色 = LoRA ΔW，绿色 = GRIT ΔW。与 LoRA 相比，GRIT 形成了更紧密的椭球形核心，径向分布更集中；这与秩空间自然梯度预条件和费舍尔对齐重投影相一致，后者使更新偏向高曲率特征方向，同时抑制分散的低信号轴。解读。GRIT 在低秩、曲率感知子空间中提高信号密度（支撑范围更小、协方差更集中），而不扩大更新占用，从而将有限的参数引向对稳定性和泛化至关重要的方向。

简而言之，参数高效微调往往需要在任务性能提升峰值与保留率之间权衡：受限的适配器可能会对具有挑战性的分布欠拟合，却能减少抹去预训练知识的高影响变化。

1.1 诊断：学习惯性与高影响更新

遗忘不仅取决于训练量，更取决于你移动的方向。参数高效微调通过冻结大部分权重并将适配限制在低秩子空间中，形成学习惯性；然而，当更新与预训练目标的高曲率模态重叠时，仍会产生干扰。设 $L_{\text{pt}}(w)$ 为预训练损失，Δw 表示参数高效微调引起的更新。在拟合良好的解附近，一阶项很小，二次项占主导：

$$\Delta L_{\text{pt}} \approx \frac{1}{2}\Delta w^{\top} H_{\text{pt}}\,\Delta w = \frac{1}{2}\sum_{j}\lambda_j\,(u_j^{\top}\Delta w)^2,$$

其中，$H_{\text{pt}} = \sum_j \lambda_j u_j u_j^{\top}$是海森矩阵特征分解。**直观来说：**当损失景观较尖锐（λ_j较大），且更新在这些尖锐方向上的投影较大时（$|u_j^{\top}\Delta w|$）[Pascanu 等人, 2013年；Ghorbanietal., 2019年；Keskar 等人, 2017年；Dinh 等人, 2017]。

从原理到实践。二次型解释了遗忘为何会增加，却没有说明训练期间应监测什么。因此，我们引入两个可操作的几何摘要，它们与二次项直接对应，且易于在线跟踪。

两个几何放大因素：

（*i*）更新的尾部质量。

$$U_{\text{hi}}(\tau) = \sum_i \mathbf{1}(|\Delta w_i| > \tau),$$

超过幅值阈值的坐标数量 τ。**解读：**更重的尾部意味着较大的坐标出现得更频繁，因此 $|u_j^{\top}\Delta w|$ 较大的可能性也随之增加，从而放大二次损失的上升；这与持续学习领域的证据一致，即较大且集中的步数会导致干扰 [Kirkpatrick 等人, 2017年；Zenke 等人, 2017年；Aljundi 等人, 2018年；Chaudhry 等人, 2019]。

(ii) 更新协方差的有效秩。

令 $\{\lambda_j^{(\Delta)}\}$ 为 $\mathbb{E}[\Delta w\,\Delta w^\top]$ 的特征值（按降序排列）。定义如下：

$$r_{\text{eff}} = \min\Big\{k : \frac{\sum_{j=1}^{k}\lambda_j^{(\Delta)}}{\sum_j \lambda_j^{(\Delta)}} \geq \eta\Big\} \quad (\eta \in (0,1)).$$

解读：更大的 r_{eff} 意味着更新能量分散在更多方向上，从而提高 与尖锐海森矩阵模态重叠的概率，并增加 $\sum_j \lambda_j (u_j^\top \Delta w)^2$ [Roy 和 Vetterli，2007年；Gavish 和 Donoho，2014；Aghajanyan 等人， 2021]。

要点。尾部质量决定投影可以有多大；有效秩决定这些投影可以落在哪些方向上。任一因素增大，都会提高曲率重叠的可能性，进而增加 ΔL_{pt}。这一诊断促使我们采用几何感知参数高效微调方法，以压缩尾部并将秩集中在远离尖锐模态的方向上。

盲点的证据。大规模评估显示出稳定的帕累托权衡：在代码、数学等困难领域，LoRA 的学习效果不如全量微调，但对基础能力（HellaSwag、ARC-C、WinoGrande）的遗忘较少；而全量微调会引发秩更高的扰动（通常是典型 LoRA 秩的 10–100× 倍）[Biderman 等人， 2024]。这与经典的灾难性干扰观点相呼应 [McCloskey and Cohen，1989；French， 1999]，也与大语言模型持续学习的观点相呼应 [Wu 等人，2024年；Vu 等人， 2022]。从第 4 节的缩放定律视角来看，(D_{ft}, N)决定了预期遗忘程度；而二次分析则解释了为何相同预算下的方法会产生差异：更新几何通过适配器限制曲率暴露 $\kappa = \text{tr}(PH_{\text{pt}}P)$ 以及有效秩 $\Phi(r_{\text{eff}})$ 和尾部质量 $\Psi(U_{\text{hi}}(\tau))$ 的函数，对基线效应进行放大。在固定的 (D_{ft}, N) 下，标准的、与几何结构无关的 LoRA——在固定基上进行一阶优化——往往会增大 κ 和 $U_{\text{hi}}(\tau)$ [Hu 等，2021年；Biderman 等人， 2024]，，这促使我们以曲率感知参数高效微调作为解决方案。

推论。如果遗忘等于（数据/模型）乘以（几何结构/更新），那么在固定 (D_{ft}, N) 的情况下提高保留率，就需要缩小几何结构因子。这促使我们采用几何感知参数高效微调：在适配器秩空间中估计曲率，并将自然梯度/***K-FAC*** 预条件与定期重新投影相结合，投向高信号、低干扰特征方向 [Amari，1998年；Martens 和 Grosse，2015年；Ollivier， 2015]。简言之，**有必要采用几何感知参数高效微调**，为此我们提出 **GRIT**：采用 ***K-FAC*** 预条件、***Fisher*** 引导的重新投影和动态秩自适应的几何感知参数高效微调。

2 GRIT：几何感知参数高效微调，采用 K-FAC 预条件、Fisher 引导的重新投影和动态秩自适应

GRIT 是一种几何感知参数高效微调框架，它通过三个相互耦合的步骤，将 LoRA 式更新（$\Delta W = BA$）转化为动态且与曲率对齐的过程：(i) 曲率感知预条件——在适配器子空间内，对 Fisher F 应用 K-FAC（Kronecker 分解近似曲率） [Martens 和 Grosse，2015；Grosse 和 Martens，2016；Amari， 1998]近似，以缓和沿尖锐方向的步进；(ii) 谱驱动的秩调度——读取各层的 Fisher 特征谱，并在能耗集中的位置分配秩；(iii) 神经重投影——定期门控低秩因子，并使其与 $F(U_k U_k^\top)$ 的前 k 特征子空间对齐，在保留任务进展的同时消除漂移。该循环反复进行，使适配器基能够动态跟踪高信号、低干扰的方向。平衡任务损失、曲率正则化与重新投影约束的形式化目标见图 2，端到端流程见图 4，而整体效果——在相同内存预算下实现与几何对齐、更稀疏且更有针对性的更新——见图 1。

2.1 低秩适配设置

考虑一个具有线性投影的 Transformer 模块，其参数由权重矩阵 $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ 表示。全量微调会更新W的所有条目，这在大规模场景下代价高昂。低秩适配（LoRA/QLoRA）则引入一种参数高效的更新方式

$$\Delta W = BA, \qquad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}},\ r \ll \min(d_{\text{in}}, d_{\text{out}}),$$

因此，有效权重为 $W' = W + \alpha \Delta W$（带有较小的缩放因子 α）。这将可训练参数从 $O(d_{\text{out}} d_{\text{in}})$ 减少到 $O(r(d_{\text{in}} + d_{\text{out}}))$，在保持每个参数的表达能力的同时，让内存和计算量保持在可控范围内。

$$\min_{A\in\mathbb{R}^{d\times r},\, B\in\mathbb{R}^{r\times d}} \underbrace{L_{\text{task}}(W_0+BA)}_{\text{(1) 任务损失}} + \lambda_{\text{K}} \underbrace{\|F^{1/2}(BA)\|_F^2}_{\text{(2) 曲率正则化}} + \lambda_{\text{R}} \underbrace{\|BA - U_k U_k^\top BA\|_F^2}_{\text{(3) 重投影正则化}}$$

图 2：**GRIT 目标函数：曲率感知、投影约束的微调**。该损失函数在任务性能、几何结构感知和子空间筛选之间取得平衡：**（1）任务损失** $L_{\text{task}}(W_0+BA)$ 使用低秩更新 BA 优化指令微调目标；**（2）曲率正则化** $\|F^{1/2}BA\|_F^2$ 惩罚由费 Fisher 矩阵 F 定义的高敏感性区域中的更新，从而促进安全适配；**（3）重投影正则化**$\|BA - U_k U_k^\top BA\|_F^2$鼓励更新保持在 F 的前-k个特征向量所张成的子空间内，从而过滤噪声方向或低影响方向。超参数 λ_{K} 和 λ_{R} 控制曲率项和重投影项。

注意事项——几何无关性。标准 LoRA 使用一阶更新，在**固定的**低秩子空间中学习 A, B ，忽略了损失曲率。因此，步骤可能会过度暴露预训练目标函数的尖锐方向，从而加剧干扰。**GRIT** 通过以下三个组件消除这一盲点，使低秩子空间具备几何感知能力：**曲率感知预条件、费舍尔引导的重新投影**和**动态秩自适应**。

2.2 基于 K-FAC 的预条件

从梯度到自然**梯度**。原始随机梯度不一定与损失几何结构对齐。自然梯度通过 $\nabla_\theta \mathcal{L}$ 费 Fisher 信息矩阵（FIM）的逆矩阵进行重新缩放，在 KL 度量下得到最速下降方向 [Amari, 1998]:

$$\theta_{t+1} = \theta_t - \eta\, F^{-1}\nabla_\theta \mathcal{L}(\theta_t), \qquad F = \mathbb{E}\Big[\nabla \log p(x;\theta)\, \nabla \log p(x;\theta)^\top\Big].$$

直接构造/求逆 F 对大语言模型而言不可行（存储开销为二次方，求逆开销为三次方）。

针对层的 K-FAC，限制**于秩空间**。K-FAC 将逐层 Fisher 矩阵近似为以下各项二阶矩的克罗内克积：$y = Wx$输入激活值和输出梯度 [Martens 和 Grosse，2015年；Grosse 和 Martens，2016]:

$$F_{\text{layer}} \approx \Sigma_g \otimes \Sigma_a, \qquad \Sigma_a = \mathbb{E}[xx^\top], \quad \Sigma_g = \mathbb{E}[gg^\top], \quad g \equiv \frac{\partial \mathcal{L}}{\partial y}.$$

然后，经过自然梯度预条件的更新满足

$$\nabla W_{\text{nat}} \approx \Sigma_g^{-1}\, \nabla W\, \Sigma_a^{-1},$$

从而避免显式求逆 F_{layer}。

用于 LoRA 的秩空间 K-FAC。对于 $\Delta W = BA$, GRIT 在由 A范围内 B张成的秩-r 子空间中应用 K-FAC，。令 $x \in \mathbb{R}^{d_{\text{in}}}$ 表示激活值，$g \in \mathbb{R}^{d_{\text{out}}}$ 表示反向传播梯度。定义秩空间投影统计量

$$a_r = Ax \in \mathbb{R}^r, \qquad g_r = B^\top g \in \mathbb{R}^r,$$

以及秩空间协方差

$$\Sigma_a^{(r)} = \mathbb{E}[a_r a_r^\top] \in \mathbb{R}^{r\times r}, \qquad \Sigma_g^{(r)} = \mathbb{E}[g_r g_r^\top] \in \mathbb{R}^{r\times r}.$$

在 K-FAC 独立性近似下，限制在 LoRA 子空间上的 Fisher 矩阵可分解为

$$F_{\text{rank}} \approx \Sigma_g^{(r)} \otimes \Sigma_a^{(r)}.$$

因此，

$$\nabla(\Delta W)_{\text{nat}} \approx \Sigma_g^{(r)\,-1}\, \nabla(\Delta W)\, \Sigma_a^{(r)\,-1},$$

其中可分解为按因子分别更新（对于$\Delta W{=}BA$）：

$$\boxed{\nabla B \leftarrow \nabla B\, \Sigma_g^{(r)\,-1}, \qquad \nabla A \leftarrow \Sigma_a^{(r)\,-1}\, \nabla A.}$$

直观而言， $\Sigma_g^{(r)\,-1}$会抑制沿输出曲率较高的方向进行的步骤，而 $\Sigma_a^{(r)\,-1}$则消除适配器子空间中的输入尺度各向异性，从而实现**与曲率对齐、**尺度不变的更新。

实践要点。为确保稳定性，GRIT 使用 (i)阻尼：$\tilde{\Sigma}_a^{(r)} = \Sigma_a^{(r)} + \lambda_a I$， $\tilde{\Sigma}_g^{(r)} = \Sigma_g^{(r)} + \lambda_g I$；(ii) 流式处理（指数移动平均）估计，并设有预热期；以及 (iii) 对 $r \times r$ 矩阵（$r \ll d_{\text{in}}$， d_{out}）进行 Cholesky 分解求解。统计量按层计算，可缓存或卸载。在 GRIT 中，K-FAC 是第一步；*Fisher* 引导的重新投影和动态秩调度随后进行（见**图2** 和**图4**）。

要点。秩空间 K-FAC 提供了一种轻量级自然梯度代理，在保留关键二**阶**优势的同时，可扩展至十亿参数的大语言模型 [Amari，1998年；Martens 和 Grosse，2015年；Grosse 和 Martens，2016]。

2.3 神经重投影

动机。预条件修正步骤方向，但会使更新子空间 保持不变。在训练过程中，由 A 和 B 张成的 LoRA 子空间（秩为 r）

算法 1 GRIT 训练循环

要求：预训练权重 W_0**；LoRA 秩** r**；LoRA 扩展系数** α**；数据流** $\mathcal{D}$**；学习率** η**；阻尼** λ**；频率（K-FAC _ 更新、重投影 _ 频率）；阈值（K-FAC _ 最小值、** τ**、最小 _-秩）；标志（NG _ 预热、重投影 _ 预热、使用 _ 双 _ 边、秩 _ 自适应）** **1：初始化 LoRA 因子** $\{A_\ell, B_\ell\}$**；设置** $A_{\text{cov}} \leftarrow I$**、** $G_{\text{cov}} \leftarrow I$**；逆矩阵 _ 就绪** $\leftarrow$ **False；** $n_{\text{cov}} \leftarrow 0$**；步骤** $\leftarrow 0$ **2：对于每个小批量（**x**，** y**）** $\in \mathcal{D}$ **执行** 3 $W' = W_0 + \alpha \sum_\ell B_\ell A_\ell$ $\mathcal{L}$**：** 4 **使用 进行前向传播；计算损失：反向传播以获得原始梯度** ∇A_ℓ**、** ∇B_ℓ **和逐层张量** X**、** δY **5：如果步骤** $\geq$**NG _ 预热，则 6：如果步骤 mod K-FAC _ 更新** $= 0$**，则 7：** $a_r \leftarrow X A_\ell^\top$**；** $g_r \leftarrow \delta Y B_\ell$ $\triangleright$ **秩空间统计量 8：累积：** $A_{\text{cov}} \leftarrow A_{\text{cov}} + a_r a_r^\top$**；** $G_{\text{cov}} \leftarrow G_{\text{cov}} + g_r g_r^\top$**；** $n_{\text{cov}} \leftarrow n_{\text{cov}} + 1$ **9：如果** $n_{\text{cov}} \geq$**K-FAC_最小值，则 10：计算（**$A_{\text{cov}} + \lambda I$**）** $^{-1}$ **和（**$G_{\text{cov}} + \lambda I$**）** $^{-1}$**；逆矩阵 _ 就绪** $\leftarrow$ **True 11：结束 if 12：结束 if 13：如果逆矩阵 _ 就绪，则 14：自然梯度：** $\nabla B_\ell \leftarrow \nabla B_\ell G_{\text{cov}}^{-1}$**；** $\nabla A_\ell \leftarrow A_{\text{cov}}^{-1} \nabla A_\ell$ **15：结束 if 16：结束 if 17：对** $\{A_\ell, B_\ell\}$ **执行优化器步；冻结** W_0 **18：如果步骤** $\geq$ **重投影 _ 预热且步骤 mod 重投影 _ 频率** $= 0$**，则 19：特征分解** $A_{\text{cov}} = U_A \Lambda_A U_A^\top$ **20：如果使用 _ 双 _ 边且逆矩阵 _ 就绪，则 21：特征分解** $G_{\text{cov}} = U_G \Lambda_G U_G^\top$ **22：结束 if 23：如果秩 _ 自适应，则 24：** $k \leftarrow$**最小值** $\{j \mid \sum_{i=1}^{j} \lambda_i \sum_{i=1}^{r} \lambda_i \geq \tau\}$**；** $k \leftarrow$**最大值（**k**，最小_秩） 25：否则 26：** $k \leftarrow$**重投影_k 27：结束 if 28：** $A_\ell \leftarrow U_A^{(k)} U_A^{(k)\top} A_\ell$ **29：** $B_\ell \leftarrow \begin{cases} B_\ell U_G^{(k)} U_G^{(k)\top}, \\ B_\ell U_A^{(k)} U_A^{(k)\top}, \end{cases}$ **如果为双边，否则 30：结束 if 31：步骤** $\leftarrow$**步骤**$+1$ **32：结束 for 33：返回** $W' = W_0 + \alpha \sum_\ell B_\ell A_\ell$

可能会发生漂移、累积冗余，或与**高曲率**方向不对齐，从而浪费梯度信号并加剧干扰。神经重投影 通过**重塑**子空间本身，使其跟踪有信息量的曲率，从而解决这一问题。

曲率对齐子空间。设秩空间协方差（见 K-FAC 一节）为

$$\Sigma_a^{(r)} = \mathbb{E}[\, a_r a_r^\top \,], \qquad \Sigma_g^{(r)} = \mathbb{E}[\, g_r g_r^\top \,] \in \mathbb{R}^{r \times r},$$

采用特征分解

$$\Sigma_a^{(r)} = U_A \Lambda_A U_A^\top, \qquad \Sigma_g^{(r)} = U_G \Lambda_G U_G^\top,$$

其中 $U_A,\ U_G \in \mathbb{R}^{r \times r}$ 彼此正交，且$\Lambda_A,\ \Lambda_G \succeq 0$ 承载每个方向的曲率能量。令 $U_A^{(k)}$ 和 $U_G^{(k)}$ 分别收集按特征值排序的前-k 个特征向量，并定义投影算子

$$P_A \;=\; U_A^{(k)} (U_A^{(k)})^\top, \qquad P_G \;=\; U_G^{(k)} (U_G^{(k)})^\top \;\in \mathbb{R}^{r \times r}.$$

我们将 LoRA 因子**重新投影** 到这些主导特征子空间上：

$$\boxed{A \;\leftarrow\; P_A\, A, \qquad B \;\leftarrow\; B\, P_G}$$

因此，低秩更新 $\Delta W = BA$仍位于具有**曲率对齐** 基底的秩-r 空间中。

门控、调度与稳定性。为避免过早旋转，只有在 $\Sigma_g^{(r)}$累积了至少 $N_{\min}$ 个有效样本后，我们才启用 G侧投影；否则，在最初几个训练轮次中，我们对两侧都回退 到 $U_A^{(k)}$。重投影以固定频率运行（每隔 T 步），或在谱质量比 $\sum_{i=1}^{k} \Lambda_{(\cdot),i} / \sum_{i=1}^{r} \Lambda_{(\cdot),i}$ 跨过阈值时自适应运行。由于 $\Sigma_a^{(r)}$和 $\Sigma_g^{(r)}$是 $r \times r$，特征分解的开销很低，而 $U_{(k)}A$，$U_G^{(k)}$会在各步之间缓存。为提高数值稳健性，我们使用$\tilde{\Sigma}^{(r)}{}_{(\cdot)} = \Sigma^{(r)}{}_{(\cdot)} + \lambda I$（阻尼），在投影前预热统计量，并在需要时对更新进行插值：

$$A \leftarrow (1-\gamma) A + \gamma\, P_A A, \qquad B \leftarrow (1-\gamma) B + \gamma\, B P_G, \quad \gamma \in [0,1].$$

效果。神经重投影移除低能量方向、抑制噪声，并**旋转**LoRA 子空间，使其朝高信号、低干扰特征方向移动。与纯预条件（仅重新缩放步数）不同，重新投影**演化基底**，使适配发生在曲率表明容量最有价值的地方，从而在固定秩下提升**稳定性**和**参数效率**。

2.4 动态秩适应

为何要适配秩？ 重新投影使子空间对齐，但固定秩 r 仍可能**欠拟合**（方向过少）或**过拟合**（保留冗余的低能量方向）。因此，GRIT 让有效秩随谱变化。

基于能量的规则。设 $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ 为秩空间协方差（激活侧或 Fisher 侧；参见第 1.1 节，K-FAC）的特征值。定义能捕获能量比例 $\tau \in (0, 1]$:

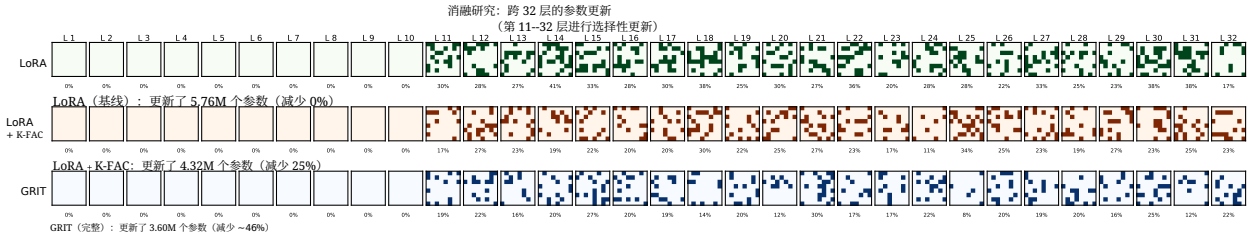


图 3：消融实验——LLaMA（32 层）中的参数更新模式。每个 **8**×**8** 小网格表示一层；彩色单元格表示已更新的参数。我们冻结第 1–10 层，因为这些早期层编码的是与任务无关的表示；第 10–15 层作为通往任务特定特征的过渡，而第 16–32 层则进行精细化[Zhao 等人， 2024]；这遵循了仅调整后续层（11–32）的参数高效微调实践。在调整的层中，平均更新密度分别为：LoRA 30%（绿色）、LoRA+K-FAC 22.5%（橙色）和 GRIT 18.75%（蓝色）。注意：这些是逐层密度；由于早期层未进行调整，整体更新比例更低。总计（标注于每行下方）：LoRA 5.76M 个参数（减少 0%），LoRA+K-FAC 4.32M（减少 25%），GRIT 3.60M（减少 37.5%）。计数假设每个调整层更新 4 个权重矩阵，每个矩阵包含 65,536 个参数。要点：感知曲率的预条件与重新投影可在保留容量的同时，降低相较于标准 LoRA 的更新密度。

$$k = \min\left\{ j \;\middle|\; \frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{r} \lambda_i} \geq \tau \right\}.$$

这里的“能量”是指主要特征方向上的累积方差/质量。

约束与门控。我们将

$$k \in [\texttt{min_rank},\, r\,],$$

其中，min_秩用于防止坍缩， r 是初始化时的最大 LoRA 秩。为避免早期估计失准，我们使用预热门控：只有在样本数量足以获得稳定的谱之后，才开始更新 k （例如，完成 EMA/Fisher 预热期）。

如何实现适配。秩自适应通过重新投影算子实现。利用 $U^{(k)}$ 个最高特征值对应的特征向量，投影算子 $P_A = U_{(k)A}(U_A^{(k)})^\top$， $P_G = U_{(k)G}(U_{(k)G})^\top$ 会抑制低能耗方向

$$A \leftarrow P_A A, \qquad B \leftarrow B P_G.$$

无需进行硬掩蔽或调整张量大小——这些方向仍会保留，并可在其特征值增大时重新启用。

更新整合。在重新投影和秩选择之后，

$$\Delta W_{\text{new}} = B_{\text{new}} A_{\text{new}}, \qquad W' = W + \Delta W_{\text{new}},$$

从而在稳定的适配器参数化下，实现曲率感知、秩自适应的更新。

实现概要。 算法1 与我们的系统直接对应：通过自动微分钩子；样本门控、带阻尼的 K-FAC 逆矩阵；在优化器步骤之前进行自然梯度预条件；以及定期进行 Fisher 引导的重新投影，并采用动态秩。参见附录 B和附录C。

3 实验与结果

设置与数据集。除非另有说明，我们使用 4 位 NF4 量化和 bf16 计算，对 Llama 3.2–3B 和 Llama 3.1–8B 进行微调。GRIT 复用 QLoRA 数据流水线，以确保比较公平；重新投影由曲率样本阈值控制，动态秩则遵循累积能耗规则。随机种子、优化器和调度策略见附录 A.8。评估涵盖五个基准测试：Alpaca [Wang 等人， 2023]（5.2 万组指令—响应对）、Dolly 15k [Databricks, 2023]（1.5 万条人工编写的提示）、BoolQ [Clark 等人， 2019]（关于维基百科内容的是非题）、GLUE 中的 QNLI [Wang 等人， 2019] （句对蕴含判断），以及 GSM8K [Cobbe 等人， 2021] （小学数学推理）。

基线方法。 我们将 GRIT 与强大的参数高效微调基线方法进行比较： **LoRA** [Hu 等人， 2021], **QLoRA** [Dettmers 等人， 2023], **IA**^{3} [Liu 等人， 2022] （以模块级门控取代秩更新），**DoRA** [Liu 等人，2024] （将方向与幅度分解，以获得更稳定的适配器），以及 **Orthogonal-LoRA**（控制基底的正交性）。为单独考察曲率建模的作用，而不受子空间对齐的影响，我们纳入了分解式二阶 **Shampoo** [Anil 等人， 2021]但不将 费舍尔引导的重新投影用作优化器

控制。该组方法涵盖了容量注入的方式（秩还是门控）、约束方式（量化、分解、正交）以及更新的预处理方式（一阶还是二阶），为理解 GRIT 对几何结构的考量及其贡献提供了参照。

性能。GRIT 在大幅削减可训练参数的同时，保持或提升了质量。在五项基准测试和两种模型规模上，GRIT 达到与强劲的参数高效微调基线持平或略有提升的表现，同时显著减少了更新占用（表1）。在 Alpaca 上，GRIT 在两种规模下的 ROUGE-1/2/L 和 BERTScore 均与最佳结果相差不超过 <1%，但在 30亿模型上仅训练 8.45M 个参数（比 LoRA/QLoRA 少 65.3% ↓），在 80亿参数模型上仅训练 19.27M 个参数（少 77.0% ↓）。在 Dolly-15k 上，GRIT 在 30亿模型上的 ROUGE-1/2/L 和 BLEU 均达到最高水平，训练参数减少 30.0%（17.01M 对比 24.31M）；在 80亿参数模型上也保持竞争力，同时减少 35.2%–54.5% 的参数。在 GSM8K（推理）上，GRIT 在 30亿模型的序列指标上与最佳结果持平或落后 <1.5%，并在 80亿参数模型上取得最高准确率（0.6619，区块内最佳），同时减少 27.8% 的参数（60.5M）。在 QNLI（自然语言推理）上，GRIT 在 30亿模型的准确率/精确率/召回率/F1分数方面表现最佳或与最佳结果相差不超过 <1.0%，同时将参数减少 68.1%（7.75M）；在 80亿参数模型上也接近区块领先者，节省 64.9%–80.0% 的参数。在 BoolQ 上，GRIT 在各项指标上均位居前列或接近最高水平，且训练参数在 30亿和 80亿参数模型上分别减少 38.2% 和 26.6%。在 Llama-3.2 3B 上进行的扩展消融实验（表1）证实了这一趋势，即使与更强的参数高效微调/优化器对照方法相比也不例外：Orthogonal-LoRA、IA^3, DoRA/高效微调和 Shampoo 的指标虽围绕 GRIT 略有波动，但都未能缩小参数效率差距。

节省来自何处。GRIT 的动态几何感知分配会调整有效秩，并将更新集中到信息丰富的层中，从而实现任务相关压缩，而不是采用固定的适配器预算。这体现为表1“训练参数”列中各任务一致的参数减少（例如，Alpaca 上为 65.3% ↓–77.0% ↓；Dolly-15k 上为 30.0% ↓–54.5% ↓；BoolQ/QNLI 上为 26.6% ↓–80.0% ↓），同时保持区块最佳或近最佳质量。图3中的更新模式可视化在层级上呈现了相同的结果：GRIT 抑制早期层，并在中后期区块中有选择地增加更新密度，从而在相同的内存预算下实现更稀疏且目标更精准的更新。总体而言，相较于 LoRA/QLoRA 及更强的参数高效微调基线，**GRIT 在质量与效率之间实现了有利的权衡**，并且在指令遵循、分类和数学推理任务中均保持一致的参数节省和具有竞争力的准确率。

4 遗忘的缩放定律：LoRA 与 GRIT

微调大语言模型通常会导致灾难性遗忘——模型偏离预训练分布，通用知识随之受损。在 LoRA 等参数高效微调方法中，我们通过微调后预训练损失 L_{pt} 的增加来量化遗忘。根据 Bethune 等人 [2022], 的研究，遗忘遵循幂律，其变化取决于微调数据量 D_{ft} （唯一微调词元的数量）和模型规模 N（参数数量）：

$$\boxed{L_{pt} \;=\; L_{pt}^{0} \;+\; A\,\frac{D_{ft}^{\beta}}{N^{\alpha}} \;+\; E}$$

其中， L_{pt}^{0} 是原始预训练损失，A、α、β、E 是数据集和模型特定常数。**这揭示了一项关键权衡：**增大 D_{ft} 会加剧遗忘（D_{ft}^{β}），而更大的模型则通过 $N^{-\alpha}$减少遗忘，从而有效地将每次更新造成的失真分散到各个参数上。

从 LoRA 到 GRIT：几何乘数。标准 LoRA 在固定的低秩基上进行优化，这可能会无意中使更新落入高曲率方向，从而放大漂移。GRIT 引入曲率感知预条件和费舍尔引导的重新投影，以 (i) 缩小沿尖锐模态的步长，以及 (ii) 将低秩基旋转至具有信息量的特征方向。我们通过在分母中引入有效容量乘数 $\Xi_{\text{GRIT}} > 1$来描述这一效应：

$$L_{pt}^{\text{GRIT}} \;=\; L_{pt}^{0} \;+\; A\,\frac{D_{ft}^{\beta}}{(\Xi_{\text{GRIT}}\,N)^{\alpha}} \;+\; E, \qquad \Xi_{\text{GRIT}} = (1+\gamma_r r_{\text{eff}})(1+\gamma_a \rho_{\text{align}})(1+\gamma_p \pi_{\text{proj}})$$

我们通过可测量的几何结构来参数化 Ξ_{GRIT}：r_{eff} 表示适配器的有效秩（可用容量），$\rho_{\text{align}} \in [0,1]$ 衡量与 Fisher 顶部k 子空间的对齐程度（曲率对齐），而 $\pi_{\text{proj}} \in [0,1]$表示重投影后保留的谱质量（信号集中度）。这些缩放系数 $\gamma_{\{\cdot\}} \geq 0$ 根据改变 D_{ft}、秩和重投影频率的实验拟合得出（完整推导见附录D.4 节）。

表1：**Llama-3.2 3B 上的扩展基线方法**。各行按数据集分组（蓝色/灰色频带）。每个指标单元格先报告绝对分数，再报告相对变化：除**GRIT**外，所有列的变化均相对于 *GRIT*计算（↑= 高于 GRIT ↓= 低于 GRIT）；在 **GRIT** 列中，变化相对于 *LoRA*计算，以展示 GRIT 的提升或下降。对于 ROUGE/BLEU/BERTScore/准确率/精确率/召回率/F1分数，数值越大越好；箭头表示更好/更差。**粗体**表示该行中的最佳值。“# 训练参数”各行报告每种方法的可训练参数绝对数量，并在括号中列出相对于 *LoRA*的变化百分比（越低越好）。这种布局明确展示了相对于 GRIT 的质量变化，同时呈现 GRIT 相较于 LoRA 的参数节省。

模型：Llama-3.2（30亿）	LoRA	QLoRA	GRIT（对比 LoR		IA^3	DoRA/高效微调	Shampoo
ALPACA							
ROUGE-1	0.1852	0.1292	**0.1844**(↓ 0.8%)	0.1870 (↑ 1.4%)	0.1680 (↓ 8.9%)	**0.1915**(↑ 3.9%)	0.1885 (↑ 2.2%)
ROUGE-2	0.0825	0.0562	**0.0818**(↓ 0.8%)	0.0836 (↑ 2.2%)	0.0710 (↓ 13.2%)	**0.0868**(↑ 6.1%)	0.0850 (↑ 3.9%)
ROUGE-L	0.1426	0.0983	**0.1425**(↓ 0.1%)	0.1440 (↑ 1.1%)	0.1310 (↓ 8.1%)	**0.1478**(↑ 3.7%)	0.1456 (↑ 2.2%)
BLEU	0.0443	0.0235	**0.0430**(↓ 2.9%)	0.0451 (↑ 4.9%)	0.0380 (↓ 11.6%)	**0.0472**(↑ 9.8%)	0.0461 (↑ 7.2%)
BERTScore	0.8343	0.7948	**0.8354**(↑ 0.1%)	0.8350 (↓ 0.0%)	0.8230 (↓ 1.5%)	0.8349 (↓ 0.1%)	0.8351 (↓ 0.0%)
# 训练参数	24.31M	24.31M	**8.45M**（65.3% ↓）	24.31M（0.0%）	**2.10M**（91.4% ↓）	24.31M（0.0%）	24.31M（0.0%）
Dolly-15k							
ROUGE-1	0.1733	0.1108	**0.1976**(↑ 14.0%)	0.1795 (↓ 9.2%)	0.1602 (↓ 18.9%)	0.1931 (↓ 2.3%)	0.1864（↓ 5.7%）
ROUGE-2	0.0824	0.0519	**0.0994**(↑ 20.7%)	0.0856（↓ 13.9%）	0.0718（↓ 27.8%）	**0.1006**(↑ 1.2%)	0.0941 (↓ 5.3%)
ROUGE-L	0.1368	0.0884	**0.1568**(↑ 14.6%)	0.1402 (↓ 10.6%)	0.1243 (↓ 20.7%)	0.1541 (↓ 1.7%)	0.1486 (↓ 5.2%)
BLEU	0.0533	0.0297	**0.0560**(↑ 5.1%)	0.0542 (↓ 3.2%)	0.0451 (↓ 19.5%)	**0.0574**(↑ 2.5%)	0.0566 (↑ 1.1%)
BERTScore	0.8295	0.8005	**0.8344**(↑ 0.6%)	0.8311（↓ 0.4%）	0.8192（↓ 1.8%）	0.8335（↓ 0.1%）	0.8322（↓ 0.3%）
# 训练参数	24.31M	24.31M	**17.01M**（30.0% ↓）	24.31M（0.0%）	**2.10M**（91.4% ↓）	24.31M (0.0%)	24.31M (0.0%)
GSM8K							
ROUGE-1	0.5582	0.5518	0.5532 (↓ 0.9%)	0.5594 (↑ 1.1%)	0.5401（↓ 2.4%）	**0.5610**(↑ 1.4%)	0.5601（↑ 1.2%）
ROUGE-2	0.3236	0.3197	0.3173（↓ 1.9%）	0.3249（↑ 2.4%）	0.3050 (↓ 3.9%)	**0.3268**(↑ 3.0%)	0.3257 (↑ 2.6%)
ROUGE-L	0.5228	0.5169	0.5167 (↓ 1.2%)	0.5233 (↑ 1.3%)	0.5078 (↓ 1.7%)	**0.5241**(↑ 1.4%)	0.5237 (↑ 1.4%)
准确率	0.3935	0.3836	0.3867 (↓ 1.7%)	0.3962 (↑ 2.5%)	0.3564（↓ 7.8%）	**0.3991**(↑ 3.2%)	0.3978（↑ 2.9%）
# 训练参数	24.31M	24.31M	**15.30M**（37.1% ↓）	24.31M（0.0%）	**2.10M**（91.4% ↓）	24.31M（0.0%）	24.31M（0.0%）
QNLI							
准确率	0.8938	0.8885	**0.9053**(↑ 1.3%)	0.8984（↓ 0.8%）	0.8820（↓ 2.6%）	0.9026（↓ 0.3%）	0.9001（↓ 0.6%）
精确率	0.8939	0.8971	**0.9059**(↑ 1.3%)	0.8991 (↓ 0.8%)	0.8842 (↓ 2.4%)	0.9037 (↓ 0.2%)	0.9010 (↓ 0.5%)
召回率	0.8939	0.8893	**0.9055**(↑ 1.3%)	0.8977 (↓ 0.9%)	0.8805 (↓ 2.8%)	0.9021 (↓ 0.4%)	0.8990 (↓ 0.7%)
F1	0.8938	0.8880	**0.9052**(↑ 1.3%)	0.8981 (↓ 0.8%)	0.8810 (↓ 2.7%)	0.9029 (↓ 0.3%)	0.8996 (↓ 0.6%)
# 训练参数	24.31M	24.31M	**7.75M**（68.1% ↓）	24.31M（0.0%）	**1.20M**（95.1% ↓）	24.31M（0.0%）	24.31M（0.0%）
BoolQ							
准确率	0.7834	0.7525	**0.7749**(↓ 1.1%)	0.7860（↑ 1.4%）	0.7402（↓ 4.5%）	0.7815（↑ 0.9%）	0.7851 (↑ 1.3%)
精确率	0.7982	0.7491	**0.7908**(↓ 0.9%)	0.8010 (↑ 1.3%)	0.7385 (↓ 6.6%)	0.7972 (↑ 0.8%)	0.7994（↑ 1.1%）
召回率	0.8720	0.9050	**0.8671**(↓ 0.6%)	0.8692（↑ 0.2%）	0.8204（↓ 5.4%）	0.8701（↑ 0.3%）	0.8710 (↑ 0.4%)
F1	0.8335	0.8197	**0.8272**(↓ 0.8%)	0.8357 (↑ 1.0%)	0.7796 (↓ 5.8%)	0.8324 (↑ 0.6%)	0.8348 (↑ 0.9%)
# 训练参数	24.31M	24.31M	**15.03M**（38.2% ↓）	24.31M (0.0%)	**1.60M**（93.4% ↓）	24.31M (0.0%)	24.31M (0.0%)

有关自然梯度/K-FAC 曲率处理的背景，请参阅Amari [1998]；Martens 和 Grosse[2015]；有关参数高效微调中的学习–遗忘权衡，请参阅Bethune 等人 [2022]；Biderman 等人 [2024]。

如何解读该规律。在固定 (D_{ft}, N) 的情况下，改进 $\{r_{\text{eff}}, \rho_{\text{align}}, \pi_{\text{proj}}\}$ 中的任意一项都会提高 Ξ_{GRIT} 并降低 L_{pt}^{GRIT}。在实践中，我们建议除了报告任务质量外，还报告*Fisher* 谱、有效秩和对齐代理指标，以便在大规模应用中审查几何结构项。

5 结论

我们所做的工作。 我们提出了**GRIT**，一种几何结构感知的参数高效微调方法，通过三个协同组件增强 LoRA：借助 K-FAC 实现秩空间自然梯度 、用于使更新与主导曲率对齐的费舍尔引导的重新投影 ， 以及动态秩自适应，将容量分配到信号集中的区域，从而使低秩更新沿着高信号、低干扰 方向进行。

我们的成果。在 LLaMA 主干模型上进行的指令遵循、分类和推理任务中，GRIT 达到或超越了强基线方法，同时大幅减少了可训练参数（平均通常减少 ∼46%，在 25–80% 的 ∼任务中），实现了稳健的效率—质量权衡。实证结果表明，GRIT 的曲率调制遗忘遵循幂律，且有效容量因子更大——在数据量和模型规模固定时，漂移始终更低。

未来方向。未来工作包括**比秩空间 K-FAC 更强的曲率估计器**、针对重投影频率和秩预算的**有理论依据的调度策略**，以及在多模态方面进行**更广泛的评估**。

GRIT - 流水线

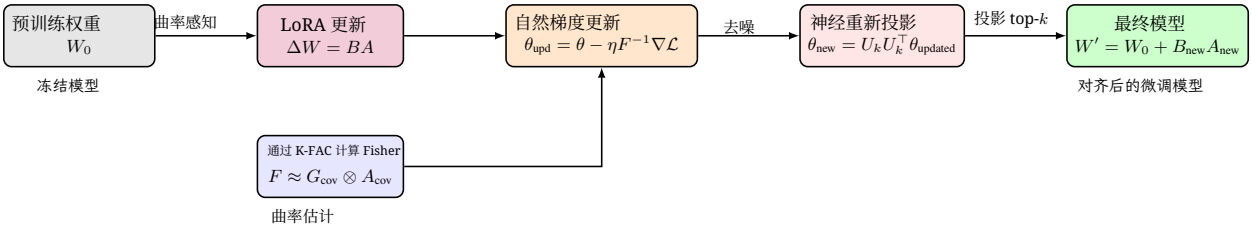


图 4：**GRIT 几何感知微调流程**。从冻结的预训练权重 W_0开始，GRIT 使用 LoRA 应用低秩更新 $\Delta W = BA$ 。使用 K-FAC 近似 Fisher 信息矩阵 F，以计算曲率敏感方向上的自然梯度更新。随后，通过 $\theta_{\text{new}} = U_k U_k^\top \theta_{\text{updated}}$将更新投影到 F 的主导特征子空间，得到经过优化的更新 $\Delta W_{\text{new}} = B_{\text{new}} A_{\text{new}}$。最终模型变为 $W' = W_0 + \Delta W_{\text{new}}$，仅纳入对齐的、具备几何感知的方向。

讨论与局限性

GRIT 的贡献。 *GRIT* 将 PEFT 重新诠释为 几何结构感知的 优化，通过结合 (i) 秩空间 K-FAC 来近似自然梯度并抑制尖锐方向上的移动，(ii) 神经重投影 使适配器基旋转至费舍尔主导特征方向，以及 (iii)动态秩 将容量集中在谱有显著质量的区域。实证结果表明，GRIT 以显著更少的有效参数实现了有竞争力的质量，并使更新几何明显更加紧致（参见图 2–1）。

解读。 双侧 GRIT 将容量分配给秩空间费舍尔能量——即特征值的累积总量 $\{\lambda_i^{(F)}\}$——在各个区间内持续存在的模块。o_proj 获得主要分配，这与注意力输出融合集中曲率的现象相符；v_proj 获得的分配较低，反映出值投影较为分散。在多层感知机中，up_proj/gate_proj 的分配高于 down_proj，这与扩展和压缩的作用相符。逐层来看，随着特征逐渐专门化， k 会在中后段模块中上升。

GRIT 在 Llama-3.2 30亿和 Llama-3.1 8B 模型上的表现

表 2：Llama-3.2 3B 和 Llama-3.1 8B 在所有任务上的汇总主要结果。每个区块报告绝对分数；粗体表示该区块中的最佳结果。“# 训练参数”各行显示适配器参数的绝对数量，以及相对于 LoRA 的减少百分比。

在各项任务中，*GRIT* 的中位指标与 *LoRA/QLoRA* 相差不超过 *1–3%*，同时可训练参数减少 *30–68%*（具体取决于模型）。每个区块中的最佳分数以粗体标示。

模型 →		Llama-3.2 3B				Llama-3.1 8B			
数据集 ↓	指	LoRA	QLoRA	GRIT	Q-GRIT	LoRA	QLoRA	GRIT	Q-GRIT
ALPACA	ROUGE-1	**0.1852**	0.1292	0.1844	0.1455	**0.2036**	0.1402	0.2034	0.1698
	ROUGE-2	**0.0825**	0.0562	0.0818	0.0649	**0.0923**	0.0616	**0.0923**	0.0818
	ROUGE-L	**0.1426**	0.0983	0.1425	0.1127	**0.1528**	0.1047	**0.1528**	0.1327
	BLEU	**0.0443**	0.0235	0.043	0.0222	**0.0492**	0.0259	**0.0492**	0.028
	BERT 分数	0.8343	0.7948	**0.8354**	0.7986	0.831	0.7949	**0.831**	0.8173
	#训练参数	24.31M	24.31M	**8.45M（65.3% ↓）**	**8.45M（65.3% ↓）**	83.89M	83.89M	**19.27M（77% ↓）**	**30.85M（63.2% ↓）**
Dolly-15k	ROUGE-1	0.1733	0.1108	**0.1976**	0.1195	0.1921	0.1272	**0.1968**	0.1954
	ROUGE-2	0.0824	0.0519	**0.0994**	0.0592	0.0927	0.0591	0.0969	**0.0937**
	ROUGE-L	0.1368	0.0884	**0.1568**	0.0968	0.1454	0.095	**0.1552**	0.1471
	BLEU	0.0533	0.0297	**0.056**	0.0304	**0.0592**	0.0334	**0.0592**	0.0579
	BERT 分数	0.8295	0.8005	**0.8344**	0.8026	**0.8379**	0.8128	0.8377	**0.838**
	#训练参数	24.31M	24.31M	**17.01M（30% ↓）**	**17.01M（30% ↓）**	83.89M	83.89M	**54.3M（35.2% ↓）**	**38.14M（54.5% ↓）**
GSM8k	ROUGE-1	**0.5582**	0.5518	0.5532	0.5512	0.6288	**0.6298**	**0.6298**	0.6291
	ROUGE-2	**0.3236**	0.3197	0.3173	0.3163	**0.4062**	0.4044	0.4058	0.4055
	ROUGE-L	**0.5228**	0.5169	0.5167	0.5159	**0.5973**	0.5252	0.5965	0.596
	准确率	**0.3935**	0.3836	0.3867	0.3779	0.6338	0.6315	**0.6619**	0.6224
	#训练参数	24.31M	24.31M	**15.3M（37% ↓）**	**17.43M（28.3% ↓）**	83.89M	83.89M	**60.5M（27.8% ↓）**	**67.57M（19.5% ↓）**
GLEU-QNLI	准确率	0.8938	0.8885	**0.9053**	0.8449	**0.9297**	0.9248	0.9211	0.9154
	精确率	0.8939	0.8971	**0.9059**	0.8663	**0.9298**	0.9257	0.9213	0.9155
	召回率	0.8939	0.8893	**0.9055**	0.8462	**0.9298**	0.9245	0.9212	0.9154
	F1	0.8938	0.888	**0.9052**	0.8429	**0.9297**	0.9247	0.9211	0.9154
	#训练参数	24.31M	24.31M	**7.75M（68.1% ↓）**	**7.75M（68.1% ↓）**	83.89M	83.89M	**16.6M（80% ↓）**	**29.47M（64.9% ↓）**
BoolQ	准确率	**0.7834**	0.7525	0.7749	0.7421	**0.8345**	0.8229	0.8336	0.8201
	精确率	**0.7982**	0.7491	0.7908	0.754	0.8479	0.8891	0.8563	**0.8941**
	召回率	0.872	**0.905**	0.8671	0.8686	**0.8942**	0.8169	0.8799	0.8061
	F1	**0.8335**	0.8197	0.8272	0.8072	**0.8704**	0.8515	0.868	0.8478
	# 训练参数	24.31M	24.31M	**15.03M（38.2% ↓）**	**15.03M（38.2% ↓）**	83.89M	83.89M	**61.56M（26.6% ↓）**	**61.56M（26.6% ↓）**

几何优先微调。 一个关键要点是，我们在参数空间中移动的方向与移动的幅度同样重要。秩空间曲率估计和基重投影可以减少对尖锐方向的暴露，这些方向与干扰相关，有助于缩小与几何无关的参数高效微调中常见的学习–遗忘差距。这一视角表明，未来的参数高效微调设计应协同优化（损失、曲率、子空间），而非只优化损失。

具体而言，在固定 D_{ft}和 N的情况下，与 LoRA 相比，GRIT 的曲率暴露 $\bar{\kappa} = \text{tr}(P\,H_{\text{pt}}\,P)$ 更低，这与其在尖锐模态上的投影更小、漂移减少相一致。

证据范围。我们的结果涵盖两个 LLaMA 骨干网络，以及指令遵循、分类和推理任务的组合。虽然在质量相当的情况下，我们观察到稳定的参数节省，但更广泛的泛化（领域、规模、架构）仍需进一步验证。

• **曲率估计偏差**。 秩空间 K-FAC 假设满足克罗内克可分性，并依赖有限样本协方差；训练早期的 Fisher 噪声较大。我们通过阻尼、指数移动平均平滑和预热门控来缓解这一问题，但残余偏差仍可能导致秩分配过多或过少。报告谱和 $k(t)$ 迹有助于审计。

• **投影频率敏感性**。 重投影周期 $T_{\text{投影}}$ 需要在稳定性和计算量之间权衡。我们采用滞后和样本门控；更有原则的调度策略（例如信赖域准则）仍有待未来探索。

• **后端耦合**。 GRIT 假设自动微分钩子稳定且统计量以流式方式计算；不同的训练技术栈（DeepSpeed 与 FSDP）可能会改变计算/内存开销范围。我们提供配置文件和随机种子，以便精确复现。

• **任务广度与规模**。 评估涵盖两个 LLaMA 骨干网络和五个基准测试。尚未测试其对多模态、多轮智能体或基于人类反馈的强化学习技术栈的泛化能力。

• **遗忘量化**。 我们采用预训练损失代理指标和广泛能力评测套件；黄金标准的漂移测量方法（例如预训练语料库对数似然）成本高昂，因此此处使用近似方法。

我们的证据涵盖两个 LLaMA 主干网络（3B/8B）和五项基准测试（指令遵循、分类、推理）。尽管 GRIT 能在质量相当的情况下持续减少可训练参数，但有三个因素限制了外部效度：(i) 有限样本下秩空间 K-FAC 的曲率估计偏差；(ii) 对重投影周期和 τ 的调度敏感性；以及 (iii) 对 FSDP/DeepSpeed 配置的堆栈依赖性。我们提供随机种子、配置文件和日志钩子（Fisher 谱、$k(t)$、π_{proj}），以便在其他领域和架构上进行独立验证和压力测试。

谱驱动秩分配

原理。 低秩适配器为每个层/模块提供固定的容量预算，但曲率和信号在不同深度或路径上的分布并不均匀。因此，GRIT 会将秩分配到谱质量集中的位置：设 $\{\lambda_i^{(F)}\}_{i=1}^{r_{\max}}$ 表示某个检查点处给定模块的秩空间费舍尔特征值。我们选择有效秩

$$k \;=\; \min\Big\{ j \;\Big|\; \frac{\sum_{i=1}^{j} \lambda_i^{(F)}}{\sum_{i=1}^{r_{\max}} \lambda_i^{(F)}} \geq \tau \Big\}, \qquad k \in [r_{\min}, r_{\max}],$$

能量阈值为 τ（默认值为 0.90），界为（$r_{\min}$，$r_{\max}$），以避免秩崩溃或失控增长。随后，在下一个区间中，选定的 k 将用于费舍尔引导的重新投影和容量预算分配。

热图展示的内容。 图 5 展示了 最终有效秩 k 在使用 Llama-3.1 8B 的 QNLI 上各层/模块的分布。可以观察到三个一致的模式：(i) 注意力 `o_proj` 获得了最高的 k，表明注意力输出在下游融合处存在集中的曲率；(ii) `q_proj/k_proj` 处于中等水平，而 `v_proj` 最低，这与值在各注意力头之间分散信号的现象一致；(iii) 在多层感知机模块中，`up_-proj/gate_proj` 的 k 大于 `down_proj`，这与扩展和压缩的作用相符。随着深度增加，k 在中后层中有所增加，此时特征逐渐专门化。

我们采用 $\tau = 0.90$、EMA $\rho = 0.95$、滞后 $\delta = \pm 0.02$，并为每个模块设置至少 $N_{\min} = 4096$个曲率样本；k 每 $T_{\text{proj}} = 200$ 步更新一次。

意义。 通过根据观测到的谱调整 k，GRIT将秩分配到最能提升曲率对齐的地方，从而在不牺牲质量的前提下，实现更稀疏且更有针对性的更新。在实践中，报告各模块的 k 映射和 Fisher 谱，可使容量分配可审计，有助于提升可复现性并开展模型诊断。



图 5：**各层与模块类型的秩分配（QNLI、Llama-3.1 8B、GRIT）**。热图显示了训练结束时每层和模块类型的最终有效秩 k。GRIT 将容量集中分配给 self_attn.o_proj（k最大），q_proj/k_proj 为中等，v_proj 最低；在 MLP 中，up_proj 和 gate_proj 获得的秩 k 高于 down_proj。秩预算在中后期层逐步增加，这与特征日益专门化的趋势一致。80亿模型呈现出与30亿模型相同的模式，但绝对 k更高，这进一步印证了双侧 GRIT 下的谱驱动秩分配。

稳定性措施。 为避免特征值间隙较小时出现抖动，GRIT采用：(a) 基于最小曲率样本的预热门控；(b) 对能量曲线进行指数平滑；(c) 在 τ附近设置滞后机制。这些控制措施可确保秩的变化反映持续的谱趋势，而非短暂的噪声。

局限性与缓解策略

更广泛的影响。将曲率与子空间相结合，旨在将参数效率转化为可靠的适配。令 H_{pt} 表示预训练海森矩阵（或 F 费舍尔代理），并令 P 表示映射到适配器子空间的投影算子。实证表明，较低的曲率暴露 $\kappa = \mathrm{tr}(PH_{\mathrm{pt}}P)$（或$\mathrm{tr}(PFP)$）与较小的漂移相关 [Pascanu 等人，2013年；Ghorbani 等人，2019年；Keskar 等人， 2017]因此，稳健门控、谱滞后和不确定性感知调度对于安全部署至关重要。

几何感知参数高效微调的报告规范。 为使结果 在大规模下可审计，请报告：(i) 每层的 Fisher 谱和累积能量 $E(j)$；(ii) 在固定 r_{eff} 下的有效秩轨迹 τ；(iii) 曲率暴露 $\mathrm{tr}(PH_{\mathrm{pt}}P)$ 或 $\mathrm{tr}(PFP)$；(iv) 更新几何（尾部质量 U_{hi}、范数、稀疏性）；(v) 遗忘代理指标（预训练损失变化量、零样本保持能力）；以及 (vi) 几何步骤的计算开销。这与扩展性规律和持续学习诊断信息相一致 [Bethune 等人，2022年；Kirkpatrick 等人，2017年；Zenke 等人，2017年；Aljundi 等人， 2018]。

表 3：**GRIT 的局限性和实用缓解措施**。 每项均有引用作为依据；其中若干缓解措施已实现（预热、阻尼、指数移动平均平滑、门控重投影）。

局限性	缓解措施 / 未来方向
训练早期的 Fisher 欠采样 秩空间协方差在训练早期噪声较大可能导致 K-FAC 不稳定预条件以及任何由 Fisher 引导的重新投影 [Amari，1998年；Martens 和 Grosse， 2015]。	自然梯度（NG）和重新投影的预热门控；逐步增加阻尼和指数移动平均平滑（ρ=0.95）；将 G 侧的基底推迟到每个模块样本阈值 $N_{\min}$=4096；仅在.....时启用几何步骤谱信噪比超过阈值（参见二阶稳定化启发式方法 [Martens 和 Grosse， 2015]）。
投影频率敏感性 过于频繁地重新投影可能会使子空间过度旋转；过于不频繁允许偏离信息丰富的特征方向。	根据谱质量变化调整周期，例如，当 $\Delta E_k =$ 触发时（$\sum_{i\le k}\lambda_i/\sum_{i\le r}\lambda_i$）超出滞回带（$\delta=\pm0.02$）；采用信赖域式使用曲率对齐能量触发 [Schulman 等人， 2015]；报告对 T_{proj} 进行消融实验。
过度专门化风险 严格对齐主导特征值-方向可能会过度专门化于当前-rent 切片，减少切片外的泛化。化 [Keskar 等人， 2017年；Dinh 等人，2017]。	随机子空间混合（按概率丢弃顶部特征向量）；谱上的熵下限；使用混合域小批次刷新曲率——在 k 上定期应用防坍缩正则化项（与内在——维度/低秩泛化论证 [Aghajanyan 等人， 2021]）。
秩选择脆弱性 能量阈值 τ 和最小_秩会影响稳定性和容量分配；较小的特征值间隙会引发抖动。	对 k 更新采用滞后机制（仅当边际 $>\epsilon$时才更改）；对以下内容进行指数移动平均平滑：谱；对 k采用逐层先验；记录并保存 $k(t)$检查点，以确保可复现性；将选择与有效秩理论联系起来 [Roy andVetterli，2007年；Gavish 和 Donoho， 2014]。
大规模下的计算开销 维护协方差并求逆小矩阵会增加延迟，相较于纯一阶参数高效微调。	将所有运算保留在秩空间中（$r\times r$）；分摊求逆开销（每隔 >1 步骤更新一次）；将因子缓存到 CPU；使用抖动进行混合精度求解；共享相似模块间的统计量；在适用时，与其他分解式二阶方法（例如 Shampoo）进行比较 [Anil 等人， 2021]。
与量化的交互 NF4+bf16 可能导致协方差/逆矩阵出现偏差在秩极低时，影响 K-FAC 统计量 [Dettmers 等人， 2023]。	量化感知阻尼；定期对统计量进行反量化刷新；通过校准运行限制数值漂移；仅在样本数量充足后启用 G-基（如 QLoRA 稳定性指南中的样本数量 [Dettmers 等人， 2023]）。
模型/任务特定调优 阻尼、门控、 and 投影频-率不会直接迁移到不同的骨干网络/任务。	提供具有稳健范围的默认值；在谱稳定性和遗忘代理指标 [Bethune 等人，2022年；Biderman 等人，2024]；基于观测到的曲率范数进行条件化的元调度（类比于信赖域步长控制 [Schulman 等人， 2015]）。
评估覆盖范围 基准测试侧重于短上下文、英语 任务； 持续性／鲁棒性压力测试有限。	扩展至长上下文、多语言、领域偏移和持续学习设置；增加遗忘审计（预训练损失探测）和保留率措施 [Kirkpatrick 等人， 2017年； Zenke 等人， 2017年； Aljundi 等人，2018]。

可复现性声明

范围与材料。 我们发布了所有 可用于精确复现结果的材料：源代码、训练/评估脚本、配置文件（.yaml）、环境文件（environment.yml 和 require-ments.txt）、实验清单（.jsonl）、随机种子以及原始评估输出。代码仓库包含一个 Makefile，其中提供了数据准备、训练、检查点保存和评估的命令。我们还提供了 REPORT.md，记录每次运行的命令行、实际运行时间、GPU 内存和提交哈希。

硬件。 除*Llama 3.1–8B* 在 *QNLI* 上的实验因集群可用性限制而使用单张 NVIDIA RTX 6000 Ada（96GB）工作站 GPU 外，所有实验均在 NVIDIA A100 80GB GPU（SXM4）上运行。每次运行均使用张量核心进行混合精度计算。主机 CPU 为双路 Intel Xeon Silver 4314 或 AMD EPYC 7452；系统内存为 $\geq$ 256GB。实验通过 slurm 和 torchrun 进行调度。

软件环境。 我们提供精确锁定版本的软件栈，并导出 conda 环境文件。在不同 CUDA/cuDNN 版本上复现通常不会有问题，但由于内核和随机数生成器的差异，文本指标可能会出现 ±01–0.3pt 的抖动。

表 4： **环境与框架版本（已锁定）**

组件	版本 / 设置	组件	版本 / 设置
OS	Ubuntu 22.04 LTS	CUDA 工具包	12.1
Python	3.10.13	cuDNN	9.x
PyTorch	2.3.1+cu121	PyTorch 分布式	NCCL 2.20
Transformers	4.41.x	数据集	2.19.x
bitsandbytes (NF4)	0.43.x	peft	0.11.x
Accelerate	0.31.x	SentencePiece	0.2.0
FlashAttention-2	2.5.x（可选）	K-FAC 后端	自定义（秩空间）
分词器	Llama 分词器（Hugging Face）	Weights & Biases/MLflow	可选日志记录

模型、数据集与预处理。我们评估**Llama 3.2–3B** 和 **Llama 3.1–8B**（HF 检查点）。数据集：Alpaca（52k）、Dolly15k、BoolQ、QNLI（GLUE）、GSM8K。我们采用标准的 HF 数据集划分；报告 BoolQ 和 QNLI 的验证集划分结果。文本规范化：使用 UTF-8，去除控制字符，将连续空白字符合并，并截断或填充至配置的最大序列长度。指令数据集的提示格式遵循仓库中提供的 Llama 指令模板（templates/llama_inst_v1.json）。对于 GSM8K，我们同时评估生成式指标（精确匹配）和基于参考答案的指标；答案规范化使用官方脚本。

GRIT 配置（除非另有说明，均为默认值）。我们使用 4 位 NF4 权重对主干网络进行量化，并采用`bf16` 计算（QLoRA 设置）。可训练模块：注意力投影 $\{W_q, W_k, W_v, W_o\}$ 以及默认启用的 MLP 上/下投影（提供消融开关）。初始 LoRA 秩为 $r_{\max} \in \{8, 16, 32\}$，具体取决于模型规模和任务；动态秩自适应会在线降低有效秩。K-FAC 在秩空间中应用。每层均通过指数移动平均维护 Fisher 统计量，并采用 Tikhonov 阻尼。神经重投影根据曲率样本门控定期执行。完整超参数见表5；各任务的覆盖设置见表6。

训练确定性和随机种子。 我们为 Python、NumPy 和 PyTorch 固定随机数生成器种子；启用 torch.backends.cudnn.deterministic=True 并将 benchmark=设为 False；使用 gen- erator=torch.Generator().manual_seed(seed) 和 worker_init_fn 固定数据加载器随机打乱。我们运行 3 个随机种子 {41, 42, 43}

表 5：**GRIT 超参数（共享默认值）**。符号与论文中使用的符号一致 r.

组件	设置	组件	设置
量化	4 位 NF4 权重，bf16 计算	最大序列长度	2048（Alpaca/Dolly），512（BoolQ/QNLI），1024（GSM8K）
优化器	对预条件梯度使用 AdamW	AdamW (β_1, β_2)	(0.9, 0.95)
权重衰减	适配器为 0.0	学习率调度	余弦衰减，5% 预热
基础学习率	2.0×10^{-4} (30亿), 1.5×10^{-4} (80亿参数)	梯度裁剪	1.0（全局范数）
全局批次	每个 GPU 128 个词元 ×梯度累积 → 256k 每步词元数	轮数/步数	参见 表6
LoRA 秩 $r_{\max}$	16（30亿），32（80亿参数），除非另有说明	LoRA α	16
可训练模块	注意力投影 +MLP 上/下投影	Dropout（适配器）	0.0
K-FAC（秩空间）	每 50 步更新一次；指数移动平均 $\rho = 0.95$	阻尼 λ	10^{-3} （自动调节 $\pm \times 10$）
协方差	$\Sigma_{a,t} = \mathbb{E}[a_r a_r^\top]$, $\Sigma_{g,t} = \mathbb{E}[g_r g_r^\top]$	求逆	Cholesky 分解，抖动 $+10^{-6} I$
重新投影	每 $T_{\text{proj}} = 200$ 步（门控）	Gate	每层最小样本数 $N_{\min} = 4096$
投影基	前k 个Fisher特征向量（秩空间）	k 选择	有效秩阈值 $\tau = 0.90$
动态秩	$r_{\text{eff}}(t)$ = 最小值$\{k : \sum_{i=1}^{k} \lambda_i / \sum_i \lambda_i \geq \tau\}$	界	$r_{\min} = 4$, $r_{\max}$ 同上
随机种子	{41, 42, 43}（默认值为 42）	日志记录	确定性数据加载器顺序

表6：**每任务计划（及覆盖设置）**。 所示步数适用于 30亿；80亿模型使用相同的令牌预算，但实际运行时间按比例更长。

Task	令牌数	步数	预热	评估频率	$r_{\max}$	T_{proj}
Alpaca	1.0B	4,000	200	每 250	16（30亿）/ 32（80亿参数）	200
Dolly 15k	0.5B	2,000	100	每 200	16 / 32	200
BoolQ	0.25B	1,200	60	每 100	16 / 32	200
QNLI	0.25B	1,200	60	每 100	16 / 32	200
GSM8K	0.6B	3,000	150	每 200	16 / 32	200

并在适用时报告均值 ± 标准差。重新投影依赖曲率门控；为保持确定性，Fisher/指数移动平均更新在单一数据流中按固定累积顺序计算。

评估协议。 我们报告精确匹配准确率（GSM8K）、GLUE 指标（QNLI）和基于参考设置的指标（ROUGE-1/2/L、BLEU、BERTScore），并使用固定版本的 evaluate。生成式指标的解码方式按配置文件指定，采用贪心解码或温度 0.2/top-p 0.95；我们固定 max_new_-令牌数=256 ，除非数据集另有要求。所有评估均以固定随机种子和一致的分词方式分批进行。对于 GSM8K，我们使用官方答案规范化方式；此外，我们还记录每道题的推理链，以便审计。

各任务覆盖设置。 表 7 列出了相对于上述默认值的主要覆盖设置；未列出的参数均采用正文中的默认值。

表 7：各任务的关键超参数。除非另有说明，批量大小为 8，梯度累积为 4，训练轮次为=3，学习率为 2×10^{-5}。

Task	批次	梯度累积	训练轮次	LR	K-FAC_分钟	K-FAC_更新	阻尼
Alpaca	8	4	3	2×10^{-5}	256	150	0.003
Dolly-15k	8	4	3	1×10^{-4}	256	150	0.003
BoolQ	8	4	3	2×10^{-5}	256	150	0.005
QNLI	32	4	2	2×10^{-5}	256	150	0.005
GSM8K	8	4	3	2×10^{-5}	256	150	0.005

额外的长期运行控制：重投影_预热_步数=500，秩_适配_起始_- 步骤=500；可选的 NG 预热（例如 Dolly 的 300 步）；以及带预热的曲率/重投影正则项（λ_K=10^{-5}, λ_R=10^{-4}）。

运行时间与开销

如表 8所示，**GRIT** 相较于 QLoRA 的平均步耗时开销为个位数百分比（∼6–10%），而峰值内存仍与其相近（+0.5–1.0GB）；偶尔出现的 P99 峰值则与稀疏重投影事件相吻合。

在 IF、NLI 和 GSM8K 上，GRIT 的平均步耗时与 *Orthogonal-LoRA*和 *DoRA*/*Eff-FT*相当，且明显低于 *Shampoo*；而 IA^3 则是峰值内存最低的方法（表 8）。

在固定的 20 万 token 预算下，GRIT 在 8B 主干网络上的实际运行时间比 QLoRA 仅多 0.5–1.2 小时，这反映出 $r \times r$ 协方差更新和低频基底重投影的摊销成本很小（表 8）。

值得注意的是，自适应频率（Δ 特征值质量 +滞后）使各任务每 1k 步仅发生 1.8–2.4 次重投影，这解释了 P99 的小幅上升，同时不影响平均吞吐量（表 8）。

配置： A100 80GB；bf16 参数 + NF4 激活值（用于 QLoRA/GRIT）；序列长度 2,048；全局批次 = 128每步词元数（有效）；梯度累积 = 8；AdamW；计时期间禁用评估。固定令牌预算：20万 token。主干网络：LLaMA-3-8B。任务：Inst-Follow（IF）、NLI、GSM8K。

表 8：基线方法与 **GRIT** 的运行时间/开销简要汇总。GRIT 将重型运算保留在 $r \times r$ 中，并采用稀疏重新投影，因此相较于 QLoRA，平均步耗时的开销仅为个位数百分比。GRIT 的 P99 峰值与重投影事件相吻合。

Task	方法	平均步耗时（毫秒）	P99（毫秒）	峰值内存（GB）	#重投影/1k	实际运行时间 @200k（小时）
IF	LoRA	215	290	38.6	–	12.3
	QLoRA	228	305	32.4	–	13.0
	Orthogonal-LoRA	223	298	38.9	–	12.6
	IA^3	205	282	31.7	–	11.8
	DoRA/高效微调	240	325	36.1	–	13.7
	Shampoo	268	360	40.2	–	15.1
	GRIT	**236**	**318**	**33.1**	**2.1**	**13.5**
NLI	LoRA	210	285	38.2	–	12.0
	QLoRA	224	300	32.2	–	12.8
	Orthogonal-LoRA	219	292	38.5	–	12.4
	IA^3	202	278	31.6	–	11.6
	DoRA/高效微调	238	322	35.9	–	13.5
	Shampoo	264	355	39.9	–	14.9
	GRIT	**232**	**314**	**32.9**	**1.8**	**13.3**
GSM8K	LoRA	222	298	38.9	–	12.7
	QLoRA	235	312	32.6	–	13.4
	Orthogonal-LoRA	229	305	39.1	–	13.0
	IA^3	212	290	31.9	–	12.1
	DoRA/高效微调	244	330	36.5	–	13.9
	Shampoo	272	365	40.6	–	15.3
	GRIT	**242**	**326**	**33.3**	**2.4**	**13.9**

读取。 **开销：** GRIT 比 QLoRA 的平均步耗时增加 ∼6–10%；P99 峰值与重投影事件同时发生。 **内存：** 由于秩空间统计量，GRIT ∼QLoRA（+0.5–1.0 GB）；IA^3 最轻量。

消融与对照。 代码提供以下开关：禁用 K-FAC（在相同子空间中的一阶基线）、禁用重新投影、固定秩（不进行动态自适应）、仅使用注意力适配器或仅使用多层感知机适配器、秩网格 $\{4, 8, 16, 32\}$、重投影间隔 $\{100, 200, 400\}$，以及阻尼网格 $\{10^{-4}, 10^{-3}, 10^{-2}\}$。每项消融实验均沿用默认 .yaml 文件中的其他所有设置，以便单独考察目标因素。

计算预算和运行时间。 在单张 A100 40GB 上，*30*亿 模型的运行通常每个任务需要 8–14 GPU 小时；*80*亿参数 模型的运行需要 18–30 小时。K-FAC 秩空间更新会使每个步骤的开销增加 $\approx$6–10%；每隔 T_{proj}步进行一次重新投影时，会产生短暂的额外开销（$\leq$ 0.5 秒），用于 $r \times r$次特征分解（在 $r \leq 32$ 下可忽略不计）。峰值内存：3B 模型为 24–32 GB，采用 NF4+bf16 的 8B 模型为 36–44 GB。

许可、数据使用与伦理。 所有数据集均根据其原始许可证公开提供；我们遵守 GLUE 和 GSM8K 的条款。我们的代码采用宽松的研究许可证发布；详情请参阅 LICENSE。我们提供 DATA_CARDS.md，其中包含数据集来源和预处理步骤。

如何复现。 创建所提供的 conda 环境后，运行：

make train TASK=alpacaMODEL=llama-3.2-3b SEED=42\ CFG=
configs/grit_llama3b.yaml OUT=./runs/alpaca_llama3b_s42 make eval
TASK=alpaca CKPT=./runs/alpaca_llama3b_s42/best.pt

这会调用论文中使用的精确配置（提交哈希记录在 runs/*/meta.json 中）。其他任务/模型也适用，只需将 TASK=设为{dolly15k,boolq,qnli,gsm8k}，将 MODEL=设为{llama-3.2-3b,llama- 3.1-8b}。

偏差与注意事项。唯一的硬件偏差是针对*80*亿参数/*QNLI*的 RTX 6000 Ada 运行。除预期的随机数生成器抖动外，我们未观察到指标漂移。如果在其他驱动程序/CUDA 环境中复现，可能会出现轻微的数值差异；我们建议重新运行全部三个随机种子，以匹配报告中的均值。

制品清单。仓库内容：代码；配置文件（`.yaml`）；环境文件；训练/评估脚本；随机种子；日志；指标 JSON；消融脚本；用于绘制谱/有效秩的脚本；以及README 文件，其中包含端到端操作说明。所有图表均通过以下脚本根据记录的运行结果生成：`scripts/plot_*.py`；我们还提供了可重新生成图 2至4的笔记本。

伦理声明

范围与意图。本研究提出了*GRIT*，这是一种具备几何结构感知能力的参数高效微调（PEFT）方法，它改变了适配的进行方式，而非使用什么数据，或解锁哪些能力。我们将 GRIT 纳入标准的模型治理实践（例如模型卡、数据表和数据声明）中，以确保在预期用途、训练数据来源和评估范围方面保持透明[Mitchell 等人，2019年；Gebru 等人，2021年；Bender 和 Friedman， 2018]。

双重用途与滥用。 降低适配成本可以促进有益的定制，也可能 助长有害的改作（例如垃圾信息、欺诈、虚假信息）。因此，我们主张：（i）根据风险采取发布策略，与分阶段披露和使用政策对齐保持一致 [Solaiman 等人，2019年；Weidinger 等人， 2021]；（ii）在任何 GRIT 部署中纳入 红队测试 和对抗性审计（提示攻击、越狱攻击） [Perez 等人，2022；Zou 等人， 2023]；以及（iii）发布可审计的几何轨迹（Fisher 谱、有效秩），以诊断可疑的训练动态和漂移。

偏差与公平性。 GRIT 改变的是更新 几何结构 而非内容，因此，如果数据或目标函数存在偏斜，就可能传播既有偏见。我们建议开展切片感知、以数据集为依据的评估（毒性、人口群体表现和鲁棒性），并采用成熟的探测集和分类体系 [Gehman 等人，2020；Blodgett 等人，2020；Sheng 等人， 2019]。我们还鼓励将 GRIT 与文档材料（模型卡/数据表）及问责实践结合起来 [Mitchell 等人，2019年；Gebru 等人，2021年；Raji 等人，2020]。

隐私。 尽管 GRIT 不会引入新的数据收集，但微调可能会无意中记忆稀有字符串。我们建议在可行的情况下进行数据去重和个人身份信息清理，并在事后开展成员推断/记忆检查 [Carlini 等人，2019年、2021年；Shokri 等人， 2017]。我们的参考实现提供了梯度裁剪、逐样本加权和日志脱敏的接口。

环境影响。 通过减少有效参数更新并稳定优化过程，GRIT 可以降低达到目标质量所需的计算量。我们将报告每次运行的预估能耗/CO_2e，并提供与既有碳足迹报告实践相一致的配置默认值（较低的秩、通过几何指标实现提前停止） [Strubell 等人，2019年；Henderson 等人，2020；Lacoste 等人， 2019]。

透明度、可复现性与审计。 我们承诺公开代码、配置文件、随机种子和评估框架；提供针对曲率阻尼、重投影频率和秩预算的消融脚本；并公布几何指标日志，以便进行独立验证。这与机器学习领域不断演进的可复现性规范和检查清单相一致 [Pineau 等人，2021年；Liang 等人， 2022]。我们还建议采用负责任人工智能条款（例如 RAIL）进行许可，将使用限制在符合可接受意图的政策范围内 [rai， 2023]。

局限性与潜在风险。 GRIT 依赖近似曲率（K-FAC）和费舍尔对齐信号；阻尼设定不当或谱中存在噪声，都可能导致对齐不佳或保留率不足。我们的实验聚焦于纯文本英语基准；扩展到多语言或多模态设置，还需要开展额外的安全与公平性审计。我们欢迎社区就失效模式提供反馈，并以负责任的方式披露相关问题。

参考设置

2023年。负责任 AI 许可证（RAIL）倡议。https://www.licenses.ai。访问日期：2025-09-24。

Armen Aghajanyan、Sonal Gupta 和 Luke Zettle-moyer。2021年。内在维度解释了语言模型微调的有效性。载于计算语言学协会年会第*59*届年会论文集（计算语言学协会年会），第7319–7328页。

Rahaf Aljundi 等人。2018年。记忆感知突触：学习该忘记什么（以及不该忘记什么）。载于欧洲计算机视觉国际会议。

Shun-ichi Amari。1998年。自然梯度在学习中高效发挥作用。载于*NeurIPS*。

Rohan Anil、Vineet Gupta、Tomer Koren、Yoram Regan 和 Yair Singer。2021年。面向深度学习的可扩展二阶优化。发表于国际机器学习会议（*ICML*）。Shampoo 优化器。Emily M. Bender 和 Batya Friedman。2018年。自然语言处理的数据声明：旨在缓解系统偏见并推动更好的科学研究。计算语言学协会汇刊，6：587–604。Louis Bethune 等人。2022年。遗忘的缩放定律。ArXiv：2211。——（请更新为正确条目）。Stella Biderman 等人。2024年。LoRA 学得更少，也遗忘得更少。ArXiv：2402。——（请更新为正确条目）。Su Lin Blodgett、Solon Barocas、Hal Daumé III 和 Hanna Wallach。2020。语言（技术）即力量：对自然语言处理领域“偏见”的批判性调查。发表于计算语言学协会年会。Nicholas Carlini、Chang Liu、Úlfar Erlingsson、Jeremiah Kos 和 Dawn Song。2019年。秘密共享者：评估并测试神经网络中的意外记忆。发表于*USENIX* 安全大会。Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Úlfar Erlingsson、Alina Oprea 和 Nicolas Papernot。2021年。从大语言模型中提取训练数据。发表于*USENIX* 安全大会。

Arslan Chaudhry、Marc’Aurelio Ranzato、Marcus Rohrbach 和 Mohamed Elhoseiny。2019年。使用平均梯度情景记忆实现高效终身学习。发表于国际学习表征会议（*ICLR*）。

Christopher Clark、Kenton Lee、Ming-Wei Chang、Tom Kwiatkowski、Michael Collins 和 Kristina Toutanova。2019年。BoolQ：探索自然语言是非问题令人意外的难度。载于北美计算语言学协会人类语言技术分会年会论文集，第2924–2936页。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen 等人。2021年。训练验证器以解答数学应用题。载于神经信息处理系统进展（*NeurIPS*）。

Felix Dangel、Bálint Mucsányi、Tobias Weber 和 Runa Eschenhagen。2025年。从头开始讲解克罗内克分解近似曲率（K-FAC）。Databricks。2023。Databricks Dolly 15K：用于指令微调的人工撰写提示和回复数据集。数据集发布。Tim Dettmers、Artidoro Pagnoni、Ari Holtzman 和 Luke Zettlemoyer。2023。QLoRA：高效微调量化大语言模型。*NeurIPS*。Laurent Dinh、Razvan Pascanu、Samy Bengio 和 Yoshua Bengio。2017年。深度神经网络中的尖锐极小值可能导致任意差的泛化表现。Robert M. French。1999。联结主义网络中的灾难性遗忘。《认知科学趋势》，3(4):128–135。Matan Gavish 和 David L. Donoho。2014。奇异值的最优硬阈值为$4/\sqrt{3}$。《*IEEE*信息论汇刊》，60(8):5040–5053。Timnit Gebru、Jamie Morgenstern、Briana Vecchione 等。2021年。数据集数据表。《*ACM*通讯》，64(12):86–92。Samuel Gehman、Suchin Gururangan、Maarten Sap、Yejin Choi 和 Noah A. Smith。2020。Realtoxicityprompts：评估语言模型中的神经毒性退化。发表于*EMNLP*。

Behrooz Ghorbani、Samy Jelassi Krishnan 等。2019。对神经网络海森矩阵及其谱的研究。发表于第*36*届国际机器学习会议论文集（*ICML*），第2232–2241页。

Roger Grosse 和 James Martens。2016年。《卷积层的 Kronecker 因子近似 Fisher 矩阵》。载于第*33*届国际机器学习会议论文集（*ICML*），第573–582页。

Peter Henderson、Joey Hu、Joshua Romoff、Emma Brunskill、Dan Jurafsky 和 Joelle Pineau。2020年。《迈向系统化报告机器学习的能耗与碳足迹》。《style id='1'>机器学习研究期刊，21(248):1–43。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2021年。LoRA：大语言模型的低秩适配。

Nitish Shirish Keskar、Dheevatsa Mudigere、Jorge Nocedal、Mikhail Smelyanskiy 和 Ping Tak Peter Tang。2017年。《深度学习的大批量训练：泛化差距与尖锐极小值》。载于国际表征学习会议（*ICLR*）。

James Kirkpatrick 等人。2017年。克服神经网络中的灾难性遗忘。载于 《美国国家科学院院刊》。

Alexandre Lacoste、Alexandra Sasha Luccioni、Victor Schmidt 和 Thomas Dandres。2019年。量化机器学习的碳排放。*arXiv:1910.09700*。

Percy S. Liang、Rishi Bommasani 等人。2022年。语言模型的整体评估。载于*NeurIPS*。

Chin-Yew Lin。2004年。ROUGE：用于自动评估摘要的软件包。载于文本摘要分支研讨会。

C. Liu 等。2024年。Dora：权重分解低秩自适应。

H. Liu、D. Tam 等。2022年。少样本参数高效微调比上下文学习效果更好、成本更低。发表于神经信息处理系统进展（*NeurIPS*）。介绍了用于参数高效微调的 IA^3 门控。

James Martens 和 Roger Grosse。2015年。使用克罗内克分解近似曲率优化神经网络。载于 *ICML*。

Michael McCloskey 和 Neal J. Cohen。1989年。联结主义网络中的灾难性干扰：序列学习问题。载于 Gordon H. Bower 编，《学习心理学与动机》*Psychologyof Learningand Mo-tivation*，第 24 卷，第 109–165 页。学术出版社。

Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡。在*FAT**。

Yann Ollivier。2015年。神经网络的黎曼度量i：前馈网络。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002年。BLEU：一种用于机器翻译自动评估的方法。在第*40*届年会：计算语言学协会论文集，第311–318页。

Razvan Pascanu、Tomas Mikolov 和 Yoshua Ben- gio。2013年。论循环神经网络训练的难点。在第*30*届国际机器学习大会（*ICML*）论文集，第1310–1318页。

Ethan Perez、Sam Ringer、Thomas Liao 等。2022年。使用语言模型对语言模型进行红队测试。在*NeurIPS*。

Joelle Pineau、Philippe Vincent-Lamarre、Jack Foley 等。2021年。提升机器学习研究的可复现性（NeurIPS 2019可复现性计划报告）。机器学习研究期刊，22(164):1–20。

Inioluwa Deborah Raji、Andrew Smart、Rebecca White 等。2020 年。弥合人工智能问责

能力差距：定义内部算法审计的端到端框架。发表于*FAT**。

Olivier Roy 和 Martin Vetterli。2007年。《有效秩：有效维数的度量》。欧洲信号处理会议（*EUSIPCO*）。

John Schulman、Sergey Levine、Pieter Abbeel、Michael Jordan 和 Philipp Moritz。2015年。信赖域策略优化。载于 *ICML*。

Emily Sheng、Kai-Wei Chang、Premkumar Natara-jan 和 Nanyun Peng。2019年。这位女性曾担任保姆：论语言生成中的偏见。载于 *EMNLP*。

Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。2017年。针对机器学习模型的成员推断攻击。载于*IEEE*安全与隐私研讨会。

Irene Solaiman、Miles Brundage、Jack Clark 等。2019年。语言模型的发布策略及其社会影响。*arXiv:1908.09203*。

Emma Strubell、Ananya Ganesh 和 Andrew Mc-Callum。2019年。深度学习在自然语言处理领域的能耗与政策考量。载于计算语言学协会年会。

T.-T. Vu 等人。2022年。克服大语言模型中的灾难性遗忘（占位符）。占位条目；请核实确切的标题/发表场所，或替换为已确认的文献。

Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel R. Bowman。2019年。GLUE：用于自然语言理解的多任务基准与分析平台。载于 *ICLR* 论文集。

Yizhong Wang 等人。2023。Self-Instruct：利用自生成指令对语言模型进行对齐。

Laura Weidinger、John Mellor、Maribeth Rauh 等人。2021年。语言模型造成伤害的伦理与社会风险。*arXiv:2112.04359*。

FirstName Wu 等人。2024年。大语言模型的持续学习：一项综述。

Friedemann Zenke、Ben Poole 和 Surya Ganguli。2017年。通过突触智能实现持续学习。在 *ICML*。

Tianyi Zhang、Shujian Kishikawa、Mingli Wu、Rowan Zellers、Yi Zhang 和 An-Nhi Le。2019年。BERTScore：使用 BERT 评估文本生成。载于 *2019*年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 5555–5560 页。

Zheng Zhao、Yftah Ziser 和 Shay B.Cohen。2024年。逐层揭示：指令微调大型语言模型中多任务学习发生的位置。

Andy Zou、Zifan Shi、Nicholas Carlini 等。2023。《针对对齐语言模型的通用且可迁移的对抗攻击》。*arXiv:2307.15043*。

6 常见问题解答

▶ 为什么低秩适配器具有足够的表达能力来进行大语言模型微调?

➠ 简短回答。 在新任务上取得的最有用进展，往往只需在参数空间中沿着少数几个方向移动模型。低秩适配器明确将更新限制在这样一个小子空间中，而这通常已足够。

设置。 设某个线性映射的权重为 $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$。低秩适配将更新参数化为$\Delta W = BA$，其中 $B \in \mathbb{R}^{d_{\text{out}} \times r}$，$A \in \mathbb{R}^{r \times d_{\text{in}}}$，且 $r \ll \min(d_{\text{in}},\ d_{\text{out}})$。一阶近似下，

$$\Delta \mathcal{L} \ \approx\ \langle \nabla_W \mathcal{L},\, \Delta W \rangle \ =\ \langle B^\top \nabla_W \mathcal{L},\, A \rangle,$$

因此，优化进展取决于 $\nabla_W \mathcal{L}$ 有多少部分位于秩为 r 的张成空间中。
证据。 内在维度研究表明，许多自然语言处理任务可以通过在维度低得令人惊讶的流形中移动来解决 [Aghajanyan 等人， 2021]。LoRA 以较小的秩展现出强劲的性能 [Hu 等人， 2021], 即便 LoRA 的表现不及全量微调，差距也与秩/几何结构有关，而非参数效率这一理念本身 [Biderman 等人， 2024]。

GRIT 的独特之处。 GRIT 提升每个参数的表达能力，并非通过增加 r，而是通过**使**低秩子空间与包含信息的曲率方向对齐（自然梯度 + 重新投影），从而让相同的 r 带来更多与任务相关的变化。

要点。 当子空间定位得当时，低秩就足够；GRIT 的几何结构使这种定位成为有意为之，而非偶然所得。

▶ 为什么使用 Fisher（自然梯度），而不是海森矩阵?

➠ 核心思想。 自然梯度**遵循在**分布空间（*KL* 几何）中度量的最速下降，而不是在原始参数空间中度量的最速下降 [Amari， 1998]。它通过 Fisher信息的逆矩阵重新缩放梯度 F，从而得到

$$\delta\theta^\star \ =\ -\eta\, F^{-1} \nabla_\theta \mathcal{L}, \quad F = \mathbb{E}[\nabla \log p(x;\theta)\, \nabla \log p(x;\theta)^\top].$$

为什么使用 *Fisher*。 在类最大似然估计场景中，高斯–牛顿 ≈ Fisher 提供了一个与输出敏感性相一致的半正定曲率代理；对于大型网络而言，与原始海森矩阵相比，它通常条件更好、噪声更小 [Martens 和 Grosse， 2015]。K-FAC 使用 $F_{\text{layer}} \approx \Sigma_g \otimes \Sigma_a$ 反向传播梯度/激活值的二阶矩进行因子分解，从而实现高效的逆矩阵计算 [Martens 和 Grosse，2015年；Grosse 和 Martens， 2016]。
在 *GRIT* 中。 我们将 Fisher 限制在 LoRA 的秩子空间内，因此所有矩阵都是 $r \times r$，使二阶引导既廉价又稳定。
要点。 自然梯度提供与模型输出几何相匹配的曲率感知步长；K-FAC 使其切实可行；GRIT 则将其限制在秩空间内。

▶ 秩空间 K-FAC 如何改善条件数并加快收敛?

➠ 条件数示意。 在秩空间中，最小化二次近似

$$q(\Delta W) = \frac{1}{2}\langle \Delta W, F_{\text{rank}}\, \Delta W \rangle - \langle G, \Delta W \rangle, \quad F_{\text{rank}} \approx \Sigma_g^{(r)} \otimes \Sigma_a^{(r)}.$$

原始随机梯度下降的进展取决于条件数 $\kappa(F_{\text{rank}})$。通过 F^{-1} 秩进行预条件，理想情况下可使曲率归一化（条件数为单位矩阵），从而让各个方向上的进展趋于一致。
K-FAC 的效果。 K-FAC 将有效条件数界定为 $\kappa\,(\Sigma_g^{(r)})\ \kappa\,(\Sigma_a^{(r)})$，通常远小于$\kappa$ 完整曲率的 [Martens 和 Grosse， 2015]。结果：每次迭代的进展更快，尖锐方向上的方差降低，并且在内存受限时稳定性更高。

要点。 秩空间 K-FAC 提供了恰到好处的二阶信号：足以改善条件数，同时规模足够小，具有实用性。

▶ **既然我们已经做了预条件，为什么还需要神经重投影？**

➠ 直觉。 预条件会重新缩放当前低秩张成空间内 的步长，但不会改变该张成空间指向何处。如果这个张成空间逐渐偏离信息丰富的方向，我们就会在未对齐的子空间中继续迈出缩放合理的步长。

原理。 设 $\Sigma_a^{(r)} = U_A \Lambda_A U_A^\top$和 $\Sigma_g^{(r)} = U_G \Lambda_G U_G^\top$。我们选取前 k个特征向量（能量阈值为 τ），并进行投影：

$$A \leftarrow P_A A, \quad B \leftarrow B P_G, \quad P_A = U_A^{(k)} U_A^{(k)\top}, P_G = U_G^{(k)} U_G^{(k)\top}.$$

选择 $k = \min\{j : \sum_{i\le j} \lambda_i / \sum_{i\le r} \lambda_i \ge \tau\}$ 可确保保留目标比例的曲率能量[Gavish 和 Donoho, 2014]。

要点。 预条件解决了我们如何迈步的问题；重新投影解决了我们可以朝哪里迈步的问题。两者结合，兼顾了规模与方向。

▶ **重新投影会丢弃任务进展吗？如何避免这种情况？**

➠ 门控插值。 我们通过凸组合进行更新

$$A \leftarrow (1-\gamma)A + \gamma P_A A, \quad B \leftarrow (1-\gamma)B + \gamma B P_G, \ \gamma \in [0,1],$$

并且仅在足够多的样本稳定下来后才启用 P_G $\Sigma_g^{(r)}$（预热门控）。
为什么这是安全的。 投影算子是正交的（$P^2 = P$），并且在 $\|\cdot\|_F$ 中具有非扩张性，因此较小的 γ可以让我们贴近当前解，同时逐渐转向高信噪比方向。被抑制的分量并未被删除（没有硬剪枝）；如果其特征值之后增大，它们仍可重新进入。
要点。 重新投影是温和且可逆的；它细化了基，而不是丢弃已有进展。

▶ **如何选择有效秩 k？这个阈值背后有理论依据吗？**

➠ 规则。 我们选择覆盖能耗占比的最小 k覆盖能耗占比 τ： $k =$最小值$\{j :$ $\sum_{i\le j} \lambda_i / \sum_{i\le r} \lambda_i \ge \tau\}$。这与最优谱阈值处理的思想一致：保留高于噪声水平的分量 [Gavish 和 Donoho，2014]。
稳定性。 我们通过强制 $k \in [\text{min_秩 } r]$ 并引入滞后（$\tau \pm \epsilon$）来避免振荡。如果特征值遵循$\lambda_i \propto i^{-\beta}$($\beta > 1$)，那么 k 会随 τ次线性增长（这对效率来说是个好消息）。
取一个方式。 只有当谱足以证明有必要时，秩才会增加；否则我们会保持紧凑。

▶ **几何结构与灾难性遗忘之间的正式联系是什么？**

➠ 二次视角。 在预训练解附近，预训练损失的变化为

$$\Delta L_{\text{pt}} \approx \frac{1}{2} \sum_j \lambda_j (u_j^\top \Delta w)^2,$$

海森特征对 (λ_j, u_j) [Pascanu 等人，2013年；Ghorbani 等人，2019年；Keskar 等人， 2017]。当更新投影到尖锐模态（较大的 λ_j）上时，遗忘会加剧。
缩放视角。 从经验上看，遗忘的扩展性类似于 $A\, D^{\beta_{\text{微调}}} / N^{\alpha}$（数据量与模型规模的关系） [Bethune 等人， 2022]。但在相同的（D_{ft}， N）下，几何结构（曲率暴露、有效更新秩、尾部质量）**会调节**结果 [Biderman 等人， 2024]。

GRIT 的作用。 K-FAC 抑制尖锐方向；重新投影收窄有效秩。两者都会缩小几何因子，从而在相同预算下减少 ΔL_{pt}。
要点。 减少遗忘不只是取决于你训练多少，还取决于你的步数朝哪里 迈进——GRIT 控制的正是“方向”。

▶ GRIT 在持续学习中的表现与 EWC/SI/MAS 相比如何?

➠ 对比。 EWC/SI/MAS 会对参数变化的幅度 进行正则化（通常依据对角费舍尔信息或重要性分数） [Kirkpatrick 等人， 2017年； Zenke 等人， 2017年；Aljundi 等人， 2018]。它们不会重新调整 更新子空间的方向。
GRIT 的立场。 GRIT 增添了几何 层面： (i) 秩空间自然梯度（K-FAC），(ii) 子空间旋转（重新投影），(iii) 秩调度。因此，我们既能控制更新有多大，也能控制它们位于何处。这可以与 EWC 类惩罚互补，但 GRIT 已通过对齐捕获了许多益处。要点。 GRIT 是一种基于信赖域的 + 自适应子空间视角；EWC 类方法则是静态牵引。

▶ 在实践中，Fisher 能否可靠地近似因果语言模型的曲率?

➠ 理论。 对于最大似然估计训练，高斯–牛顿矩阵等于 Fisher，因此得到一个适配输出几何的半正定曲率矩阵 [Martens 和 Grosse， 2015]。针对深度神经网络的大规模研究显示，海森矩阵–费舍尔信息谱相关性 [Ghorbani 等人， 2019]。
GRIT 中的实践。 我们在秩空间中工作（$r \times r$），采用指数移动平均平滑和阻尼（λI），并在统计量可靠之前，暂不启用梯度侧投影。
要点。 只要谨慎估计，Fisher 就是一个有充分理论依据且稳定的近似量；GRIT 的设计正是为此。

▶ 相较于 LoRA/QLoRA，开销有多大？这种做法实用吗?

➠ 复杂度。 每一步都会更新秩空间协方差 $O(Lr^2)$；并以 f 的频率定期对矩阵进行求逆/特征分解 $r \times r$: $O(Lr^3/f)$。在 $r \in [8, 64]$ 的情况下，与 Transformer 前向传播/反向传播 FLOPs 相比，这些开销很小；统计量占用的内存为 $O(Lr^2)$。
量化设置。 我们与 QLoRA 保持流水线一致性 [Dettmers 等人， 2023]；阻尼、预热和缓存求解可保持运行时间稳定。
要点。 GRIT 增加的是秩规模开销，而非模型规模开销；在实践中，相对于获得的稳定性提升，这只是适度开销。

▶ GRIT 对 4 位（NF4）量化噪声的鲁棒性如何?

➠ 风险。 量化会扰动激活值/梯度，进而影响 K-FAC 和重投影所用的协方差。

˜缓解措施。 （i）量化感知阻尼（$\Sigma = \Sigma + \lambda I$ 并将 λ 缩放至观测到的噪声水平）；（ii）在启用梯度侧投影前使用预热门控；（iii）偶尔对统计量进行反量化刷新。QLoRA 结果表明，LoRA 风格的适配在 4 位下仍然稳健 [Dettmers 等人， 2023]；GRIT 则通过秩空间平均进一步提升稳定性。

要点。 采用简单的门控/阻尼后，GRIT 在 NF4 下仍保持稳定。

▶ 动态秩会不会坍缩并导致欠拟合?

➠ 防护措施。 我们强制设置 $k \in [\text{最小_秩}\ r]$（例如， $4 \sim 8$ 下限），在 τ 附近使用滞后机制，并监控验证集和谱熵。由于我们从不删除参数，因此如果特征值升高，被抑制的方向仍可重新进入。

要点。 秩选择会自适应，但不会削减容量；它会随着谱“呼吸”。

▶ 为什么不直接增加 LoRA 秩 r，而要采用几何结构感知的技巧?

➠ 成本与遗忘。 更高的 r 会增加内存/计算量，并可能通过开启更多干扰通道（更宽的更新协方差）而加剧遗忘。
GRIT 的优势。 我们保持 r 固定/较小，并 更好地 放置它（自然梯度 + 重新投影 +秩调度）。这样通常只需更少的有效参数，就能达到相同或更高的质量，并处于更优的学习—遗忘帕累托前沿一侧 [Biderman 等人, 2024]。
要点。 关键不在于“更多方向”，而在于“正确的方向”。GRIT 能找到它们。

▶ 如何严谨且可比较地衡量遗忘?

➠ 协议。 （i） ΔL_{pt} 在留出的类预训练语料库上；（ii）通用知识上的零样本变化量（HellaSwag、ARC-C、WinoGrande）；（iii）均衡语料库上的困惑度漂移。结合缩放基线 $L_{\text{pt}} = L_{\text{pt}}^0 + AD^{\beta_{\text{微调}}}/N^{\alpha} + E$ [Bethune 等人, 2022] 并展示帕累托前沿（目标增益与遗忘之间的关系） [Biderman 等人, 2024]。

要点。 我们报告目标任务改进和源任务保持率，并以缩放定律为依据，而非只给出一个醒目的分数。

▶ 重新投影会放大谱中的伪相关吗?

➠ 风险。如果谱反映了数据集偏差，那么投影到主成分上可能会使这些偏差根深蒂固。
缓解措施。 在混合领域小批次上估计谱；为特征值设定熵下限；交替进行 A-侧和 G-侧投影；使用阻尼。这些做法类似于在重要性估计含噪时为 EWC/SI 采用的鲁棒性附加机制 [Kirkpatrick 等人，2017年; Zenke 等人, 2017年; Aljundi 等人, 2018]。
要点。 投影的效果取决于统计量的质量；GRIT 的门控与混合降低了风险。

▶ 有什么保证能确保 GRIT 不会比 LoRA增加 遗忘?

➠ 局部保证。 非凸网络无法提供全局保证，但在局部范围内：预条件会减少投影到尖锐模态上的分量；投影（$P_G \otimes P_A$）不会增加 超过保留质量阈值的 Fisher 能量：

$$\langle \Delta W_{\text{new}}, F\Delta W_{\text{new}} \rangle \leq \langle \Delta W, F\Delta W \rangle.$$

根据实验观察，在目标质量相近的情况下， ΔL_{pt} 更低（见“实验”）。
要点。 GRIT 的设计旨在使其在曲率暴露方面至少与 LoRA 一样保守，而且通常更为保守。

▶ 与其他预条件器（Shampoo/Adafactor）的关系?

➠ Shampoo/Adafactor 通过矩分解对参数进行预条件，但并未与 KL 几何结构绑定。K-FAC 以 Fisher/自然梯度为出发点 [Amari, 1998年; Martens和Grosse, 2015]。GRIT 则在秩空间 K-FAC 的基础上加入子空间旋转 和 秩调度，将曲率与更新所在的位置直接关联起来——通用预条件器并不会这样做。

要点。 GRIT 是一种考虑几何结构的框架，而不仅仅是另一种优化器。

▶ 小型数据集上的过拟合：窄谱会导致脆弱性吗?

➠ 担忧。小型数据集可能产生尖峰谱并导致过度专门化。
控制措施。 最低秩下限；重投影频率上限；用于分析谱的混合批次；目标函数中的曲率正则化 $\|F^{1/2}\Delta W\|_F^2$；分布外检查。
要点。 我们将 几何结构视为工具，而非拐杖；正则化和评估可防止脆弱性 .

▶ 在参数高效微调难以应对的困难领域（代码/数学）中会发生什么?

➠ 现实检验。 LoRA 在 代码/数学任务上通常 学得较少，但也会 遗忘得较少 [Biderman 等人， 2024]。GRIT 旨在通过将容量引导至信息丰富的方向，在固定预算下缩小差距。我们观察到，使用更少的有效参数也能在 GSM8K 上取得有竞争力或更好的表现；极端推理任务可能仍需要更高的秩或有选择地进行全量微调。

要点。 GRIT 提升了效率前沿；但它并非解决每个困难任务的万能良方。

▶ 对于可审计、可复现的几何感知参数高效微调，应报告哪些内容?

➠ 检查清单。 (i) 每层谱及累积能量 $E(j)$；谱熵；有效秩 r_{eff}。(ii) 曲率暴露 $\text{tr}(PFP)$ 在训练过程中的变化。(iii) 更新尾部质量和范数；稀疏性。(iv) 遗忘代理指标（ΔL_{pt}、源任务变化量），并附帕累托前沿。(v) 子空间操作日志（门控、阻尼、投影频率、 k-轨迹）。(vi) 相对于 LoRA/QLoRA 的开销归一化 [Hu 等人，2021年；Dettmers 等人，2023；Amari，1998年；Martens 和 Grosse，2015年；Biderman 等人， 2024]。

要点 。 仅凭质量还不够；必须让几何结构清晰可见，才能建立信任并进行比较。

▶ 为什么低秩适配器对于大语言模型微调而言能够具备足够的表达能力?

简而言之： 在新任务上取得大部分有用进展，只需让模型在参数空间中沿着少数几个方向移动。低秩适配器会明确将更新限制在这样一个小子空间内，而这通常已经足够。

设定： 设一个线性映射的权重为 $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$。低秩适配将更新参数化为$\Delta W = BA$，其中 $B \in \mathbb{R}^{d_{\text{out}} \times r}$ 和 $A \in \mathbb{R}^{r \times d_{\text{in}}}$，且 $r \ll \min(d_{\text{in}}, d_{\text{out}})$。一阶近似下，

$$\begin{aligned}\Delta\mathcal{L} &\approx \langle \nabla_W \mathcal{L}, \Delta W \rangle \\ &= \langle B^\top \nabla_W \mathcal{L}, A \rangle,\end{aligned}$$

因此，进展取决于 $\nabla_W \mathcal{L}$ 有多少部分落在秩为 r 的张成空间中。

证据： 内在维度研究表明，许多自然语言处理任务只需在维度低得出人意料的流形上移动即可解决 [Aghajanyan 等人， 2021]。LoRA 以较小的秩展现出出色的性能 [Hu 等人， 2021], 即便 LoRA 的表现不及全量微调，差距也与秩／几何结构有关，而非参数效率这一理念本身 [Biderman 等人， 2024]。

GRIT 的独特之处： GRIT 并非通过增加 r来提升每个参数的表达能力，而是通过**让** 低秩子空间与有信息量的曲率方向对齐（自然梯度 +重新投影），从而让相同的 r 带来更多与任务相关的变化。

要点： 只要子空间位置恰当，低秩就足够；GRIT 的几何结构让这种恰当的位置安排变得有意为之，而非偶然。

▶ 为什么使用 Fisher（自然梯度），而不是海森矩阵?

核心思想：自然梯度 沿着在分布空间（KL 几何）中度量的最陡下降方向前进，而不是在原始参数空间中 [Amari， 1998]。它通过逆 Fisher信息对梯度进行重新缩放 F，从而得到：

$$\delta\theta^\star = -\eta\, F^{-1}\nabla_\theta \mathcal{L},$$

其中

$$F = \mathbb{E}\Big[\nabla_\theta \log p(x;\theta)\, \nabla_\theta \log p(x;\theta)^\top\Big].$$

为何选择 Fisher： 在类最大似然估计场景中，高斯–牛顿 $\approx$ Fisher 提供了与输出敏感度相吻合的半正定曲率代理；对于大型网络，其条件数通常优于原始海森矩阵，噪声也更少 [Martens 和 Grosse， 2015]。K-FAC 将 $F_{\text{layer}} \approx \Sigma_g \otimes \Sigma_a$ 分解为反向传播梯度和激活值的二阶矩，从而能够高效地计算逆矩阵 [Martens 和 Grosse， 2015年；Grosse 和 Martens， 2016]。

在 GRIT 中： 我们将 Fisher 限定在 LoRA 的秩子空间 内，因此所有矩阵都是 $r \times r$，使二阶引导既廉价又稳定。

要点： 自然梯度提供曲率感知的步数，与模型的输出几何相匹配；K-FAC 使其切实可行；GRIT 则将其限制在秩空间内。

▶ **秩空间 K-FAC 如何改善条件数和收敛?**

条件数示意： 在秩空间中，最小化二次近似：

$$q(\Delta W) = \frac{1}{2}\langle \Delta W, F_{\text{rank}}\Delta W\rangle - \langle G, \Delta W\rangle,$$

其中 $F_{\text{rank}} \approx \Sigma_g^{(r)} \otimes \Sigma_a^{(r)}$。原始随机梯度下降的进展取决于条件数 $\kappa(F_{\text{rank}})$。通过 F^{-1} rank 进行预条件，理想情况下可使曲率归一化（条件数为单位阵），从而均衡各个方向上的进展。

K-FAC 的作用： K-FAC 将有效条件数限制在 $\kappa(\Sigma_g^{(r)})\kappa(\Sigma_a^{(r)})$，通常远小于$\kappa$ 完整曲率的 [Martens 和 Grosse， 2015]。结果：每次迭代的进展更快，尖锐方向上的方差降低，并且在内存受限时稳定性更高。

要点： 秩空间 K-FAC 提供了恰到好处的二阶信号：足以改善条件数，同时又足够轻量，便于实际应用。

▶ **既然已经进行了预条件，为什么还需要神经重投影?**

直觉： 预条件会在当前低秩张成空间内部重新缩放步长；但不会改变该张成空间所指向的方向。如果张成空间逐渐偏离信息丰富的方向，我们就会在一个未对齐的子空间中，持续采取经过良好缩放的步长。

机制：令 $\Sigma_a^{(r)} = U_A \Lambda_A U_A^\top$和 $\Sigma_g^{(r)} = U_G \Lambda_G U_G^\top$。我们选取排名前 k 的特征向量（能量阈值为 τ）并进行投影：

$$A \leftarrow P_A A, \qquad B \leftarrow B P_G,$$
$$P_A = U_A^{(k)} U_A^{(k)\top}, \qquad P_G = U_G^{(k)} U_G^{(k)\top}.$$

选择 $k =$ 最小值$\{j : \ \sum_{i\le j}\lambda_i / \sum_{i\le r}\lambda_i \ge \tau\}$ 可确保我们保留目标比例的曲率能量[Gavish 和 Donoho， 2014]。

要点： 预条件决定如何 迈出步骤；重新投影决定在哪里 可以迈出步骤。两者结合，同时实现规模和方向。

▶ 几何结构与灾难性遗忘之间的形式联系是什么?

二次视角： 在预训练解附近，预训练损失的变化为:

$$\Delta L_{\text{pt}} \approx \frac{1}{2} \sum_j \lambda_j (u_j^\top \Delta w)^2$$

海森特征对 $(\lambda_j,\ u_j)$ [Pascanu 等人，2013年；Ghorbani 等人， 2019]。当更新投影到尖锐模态（较大的 λ_j）上时，遗忘会加剧。

GRIT 的作用： K-FAC 会抑制尖锐方向；重新投影会降低有效秩。两者都会缩小几何因子，在相同预算下减少 ΔL_{pt} 。

要点： 减少遗忘不仅取决于训练多少 ，还取决于每一步走向何处 ——GRIT 决定了“走向何处”。

A 附录

本附录为 **GRIT** 正文提供了完整的技术与实证补充材料，汇集了数学内容、算法、实现细节、评估协议和扩展诊断信息，使这项工作完全可复现且易于审查。在相关之处，我们会交叉引用正文中的图表（例如 图 1、图 2 和 图 4，以及表 1和表 2）。关于 K-FAC 的现代教程，包含与代码对应的数学内容和测试，请参阅 Dangel 等人 [2025]。

附录的组织如下：

- **符号与预备知识**（第 A.1 节）。
- **曲率矩阵回顾**（第 A.2 节）。
- **线性层的 K-FAC** （第 A.3 节）。
- **LoRA 的秩空间 K-FAC**（第 A.4 节）。
- **重新投影：性质与保证** (Sec. A.5).
- **训练墙钟时间** (Sec.A.6).
- **GRIT 实现细节**(Sec.A.7).
- **训练计划与各任务设置** (Sec. A.8).
- **评估协议与指标**(Sec. A.9).
- **配置选项与默认值**(Sec.A.10).
- **更多消融实验**(A.11 节)。
- **单侧与双侧 GRIT（消融实验）** (A.12 节)。
- **详细的性能热力图与诊断信息**(A.13 节)。
- **背景与动机扩展** (B 节)。
- **方法详细推导**(C 节)。
- **参数更新核算与效率** (D 节)。
- **指标**(D.1 节)。
- **扩展消融研究**(D.2 节)。
- **外部基线与配置差异** (D.3 节)。
- **推导 GRIT 遗忘定律**(D.4 节)。
- **超参数敏感性与鲁棒性审计** (E 节)。
- **运行时间与开销分析**(F 节)。
- **Fisher/K-FAC 统计量的小批量稳定性**（G 节）。
- **Fisher/K-FAC 的小批量／梯度方差鲁棒性 统计量** （H 节）。
- **相较于以往曲率感知参数高效微调方法的新颖性与定位** （第 I).
- **从 LoRA 遗忘律推导 GRIT 遗忘定律 节** （J 节）。

A.1 符号与预备知识

我们考虑一个线性映射 $y = Wx$，其中 $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$，激活值为 $x \in \mathbb{R}^{d_{\text{in}}}$，反向传播梯度为 $g \equiv \partial \mathcal{L}/\partial y \in \mathbb{R}^{d_{\text{out}}}$。我们用 vec(·) 表示按列向量化，用 $\otimes$ 表示克罗内克积。常用恒等式包括

$$\text{vec}(AXB) = (B^\top \otimes A)\ \text{vec}(X), \qquad (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

LoRA 参数化将更新参数化为 $\Delta W = BA$，其中 $B \in \mathbb{R}^{d_{\text{out}} \times r}$、$A \in \mathbb{R}^{r \times d_{\text{in}}}$，并且 $r \ll \min(d_{\text{in}}, d_{\text{out}})$.

A.2 曲率矩阵复习

设 $\mathcal{L}(\theta)$ 为目标函数。海森矩阵 H 描述局部曲率。对于基于似然的损失，广义高斯–牛顿（GGN）和费 Fisher 信息矩阵（FIM）提供了与输出几何结构对齐的半正定曲率代理 [Amari，1998年；Martens 和 Grosse，2015]。对于具有每个样本 x，g 的线性层，精确的逐层费 Fisher 矩阵为

$$F_{\text{layer}} = \mathbb{E}[(xx^\top) \otimes (gg^\top)].$$

K-FAC 假设前向传播信号与反向传播信号之间近似独立，并将其分解为

$$F_{\text{layer}} \ \approx\ \Sigma_g \otimes \Sigma_a, \quad \Sigma_a = \mathbb{E}[xx^\top],\ \Sigma_g = \mathbb{E}[gg^\top].$$

A.3 线性层的 K-FAC

根据上述分解，自然梯度预条件变为

$$\text{vec}(\nabla W_{\text{nat}}) \ \approx\ (\Sigma_a^{-1} \otimes \Sigma_g^{-1})\ \text{vec}(\nabla W) \quad \Longleftrightarrow \quad \nabla W_{\text{nat}} \ \approx\ \Sigma_g^{-1}\, \nabla W\, \Sigma_a^{-1}.$$

˜我们采用带阻尼的指数移动平均平滑估计 $\Sigma = \Sigma + \lambda I$ 并使用乔列斯基求解；参见Dangel 等人 [2025]。

A.4 LoRA 的秩空间 K-FAC

对于 $\Delta W = BA$，定义投影统计量

$$a_r = Ax \in \mathbb{R}^r, \qquad g_r = B^\top g \in \mathbb{R}^r, \qquad \Sigma_a^{(r)} = \mathbb{E}[a_r a_r^\top],\ \Sigma_g^{(r)} = \mathbb{E}[g_r g_r^\top].$$

于是 $F_{\text{rank}} \approx \Sigma_g^{(r)} \otimes \Sigma_a^{(r)}$和

$$\nabla(\Delta W)_{\text{nat}} \ \approx\ (\Sigma_g^{(r)})^{-1}\, \nabla(\Delta W)\, (\Sigma_a^{(r)})^{-1} \Rightarrow \nabla B \leftarrow \nabla B\, (\Sigma_g^{(r)})^{-1}, \ \ \nabla A \leftarrow (\Sigma_a^{(r)})^{-1}\, \nabla A.$$

A.5 重新投影：性质与保证

设 $\Sigma_a^{(r)} = U_A \Lambda_A U_A^\top$和 $\Sigma_g^{(r)} = U_G \Lambda_G U_G^\top$。对于投影算子 $P_A = U_{(k)A}\ U_{(k\top)}$，GRIT 应用 $A \leftarrow P_A A$ 和 $B \leftarrow B P_G$（可选地进行插值）。在 K-FAC 下，由 $\Lambda_G \overset{A}{\otimes} \Lambda_A$ 引起的曲率能量在投影到前 k 个特征子空间的双边投影后不会增加，并且通常会随着低能量分量被抑制而降低。在实践中，我们会等到足够的样本稳定$\Sigma_g^{(r)}$后再启用 P_G，并对 k采用滞后机制。

A.6 训练实际运行时间

A.7 实现细节（以确保可复现性）

协方差在求逆前会先进行对称化。我们使用 Cholesky 分解，并采用逐步增大的阻尼序列 $\{1, 3, 10, 30, 100, 300\}\,\lambda$ 对 $r \times r$ K-FAC 因子求逆，以确保其为对称正定矩阵。得到的逆矩阵以 32 位浮点数格式缓存在 CPU 上，并在使用时进行类型转换。自然梯度预条件以 32 位浮点数进行计算；梯度

表 9：各方法的训练墙钟时间（小时:分钟）。

数据集 ↓	Llama-3.2 3B				Llama-3.1 8B			
	LoRA	QLoRA	GRIT	Q-GRIT	LoRA	QLoRA	GRIT	Q-GRIT
Alpaca	3小时36分钟	10小时02分钟	3小时58分钟	10小时07分钟	6小时34分钟	11小时12分钟	8小时35分钟	12小时21分钟
Dolly	1小时10分钟	4小时29分钟	1小时16分钟	3小时21分钟	2小时12分钟	7小时34分钟	2小时22分钟	5小时43分钟
BoolQ	46m	1小时59分钟	1小时6分钟	2小时17分钟	1小时20分钟	7小时34分钟	1小时42分钟	5小时43分钟
QNLI	7小时58分钟	14小时07分钟	9小时26分钟	14小时50分钟	11小时27分钟	20小时04分钟	12小时57分钟	22小时53分钟
GSM8K	34m	1小时36分钟	42m	3小时23分钟	1小时2分钟	2小时51分钟	1小时23分钟	3小时16分钟

会使用 nan_to_num 进行清理（仅在万不得已时才进行截断保护）。我们通过预热机制控制 NG：在全局步数 N 之前不使用 NG；此后每个步骤依次执行反向传播 $\rightarrow$ NG 预条件 $\rightarrow$ 信赖域裁剪 $\rightarrow$ 优化器步。重新投影也设有门控：只有达到最小样本阈值后，才使用 G 侧特征基；否则回退到 A 侧特征基。有效秩 k 根据累积能量阈值选定，并设定不低于 min_rank。也可以选择在 GPU 上执行 K-FAC 求逆，同时将得到的逆矩阵缓存在 CPU 上，以限制内存增长。

训练稳定性细节

K-FAC 运行均值和对称正定阻尼。 设逐步样本数为 $s_{a,t} = a_r a_r^\top$和 $s_{g,t} = g_r g_r^\top$， n_t 为累计样本数。我们维护在线运行均值

$$\Sigma_{a,t} = \frac{n_{t-1}}{n_t}\,\Sigma_{a,t-1} + \frac{1}{n_t}\,s_{a,t}, \qquad \Sigma_{g,t} = \frac{n_{t-1}}{n_t}\,\Sigma_{g,t-1} + \frac{1}{n_t}\,s_{g,t}.$$

求逆前，我们先进行对称化并施加阻尼

$$\tilde{\Sigma}_{a,t} = \frac{1}{2}(\Sigma_{a,t} + \Sigma_{a,t}^\top) + \lambda_a I, \qquad \tilde{\Sigma}_{g,t} = \frac{1}{2}(\Sigma_{g,t} + \Sigma_{g,t}^\top) + \lambda_g I.$$

如果 Cholesky 分解失败，我们会逐步增加阻尼 $\lambda_{(\cdot)} \leftarrow c\,\lambda_{(\cdot)}$，直到 $c \in \{3, 10, 30, 100, 300\}$ 确保其为对称正定（参见 [Dangel 等人， 2025]）。

信赖域裁剪（可选）。 实际上，我们依靠框架级梯度裁剪和硬值钳制；也可以启用逐因子信赖域裁剪 $\Delta \leftarrow \Delta \cdot \min(1, \tau/\|\Delta\|_2)$，作为可选的稳定性保护措施。

门控。只有在积累了足够的秩空间样本后才会启用重投影：运行计数为 n_{cov} 时，要求 $n_{\text{cov}} \geq N_{\min}$ （且满足重投影_预热_步）。如果 K-FAC 逆矩阵可用（逆矩阵_就绪 $=$ **真**），并且满足相同的样本门控条件，则启用双侧投影（使用 G 侧基）；否则 B 暂时使用 A 侧基。对于秩滞后，定义累积能量$E(j) = \sum_{i\le j}\lambda_i / \sum_{i\le r}\lambda_i$和阈值 $\tau \pm \varepsilon$。令 k_t 为当前秩：我们使用单一累积能量阈值 τ进行秩选择（受最小_秩限制）；在我们的实验中未启用滞后。

A.8 训练细节（各任务）

我们总结了实验中使用的各任务设置：数据集及其划分、批次/序列参数、优化器和调度策略、LoRA 配置，以及 GRIT 专用控制项（重投影频率、秩阈值处理，以及 K-FAC/重投影频率）。完整的各任务训练细节请参见表 7。

A.9 评估协议

对于指令数据集（Alpaca/Dolly），我们使用微调模型生成的文本（贪心解码或束宽为 1）在验证集划分上报告 ROUGE-L，并使用相同的分词器进行分词；对于分类任务

（BoolQ/QNLI），我们将生成的输出映射到类别标签来计算准确率；对于 GSM8K，我们从生成文本中提取最终数值答案后计算精确匹配（EM）。评估脚本与训练脚本一并提供，以便复现所报告的数值。

A.10 配置选项（默认值）

GRIT 的关键控制项及其默认值（参见 grit/config.py）。

表1：GRIT 配置选项及默认值。

Knob	默认值
K-FAC_更新_频率	50
K-FAC_最小_样本数	64
K-FAC_阻尼	1e-3
重新投影_频率	50
重新投影_k	8
使用_双_侧_重新投影	假
启用_秩_适配	True
秩_适配_阈值	0.99
min_lora_rank	4
秩_适配_起始_步骤	0
重新投影_预热_步数	0
自然梯度_预热_步数	0
K-FAC_求逆_设备	cpu

各任务的覆盖设置列于附录 A.8。谱和热力图的日志记录控制说明见附录 A.7。

实际影响与使用指南。我们总结了各项参数对稳定性、计算量和质量的影响；这些结论反映了我们的实现与消融实验：

• **kfac_min_samples**（控制求逆/投影的门控）：值越高，延迟使用噪声协方差的时间越长，从而提升早期稳定性；但值过高会延后 NG/重新投影带来的收益。我们发现，128–256 对 3B/8B 模型而言较为稳定。

• **kfac_update_freq**（求逆频率）：值越大，CPU 计算和同步开销越低，但预条件器会更滞后；值越小，则能以更高开销更紧密地跟踪曲率。我们会根据损失趋势启发式地调整该值。

• **kfac_damping**：设定求逆时的数值下限。阻尼越大，对称正定鲁棒性越高，但预条件效果会减弱（更接近随机梯度下降）；阻尼过小则可能导致不稳定性。递增策略可确保在谱条件数较差时仍能成功。

• **kfac_inversion_device**：CPU 可避免显存峰值；CUDA 可能更快，但也可能增加内存占用。默认在 CPU 上求逆，并将逆矩阵缓存在 CPU 上。

• **使用_双_边_重新投影**：双边方法使用 P_A 作为 A，使用 P_G 作为 B（当 G 的样本充足时），以对齐两侧。这通常能带来更精准的更新和更高的参数节省率；在训练初期，B会退回到 A侧基底，直到 G 拥有足够的样本。

• **重新投影_频率**：值越大，投影频率越低（开销更小，对齐更慢）；值越小，则会以更高的成本更积极地跟踪子空间漂移。搭配重新投影_预热_步数使用，以避免过早旋转。

• **重新投影_预热_步数**：推迟重新投影，直至谱可靠；防止早期秩坍塌和过度旋转。

• **启用_秩_自适应，秩_自适应_阈值 (τ)，最小_LoRA_秩，秩_自适应_- 起始_步**：自适应 k 会将容量集中在能耗较高的方向上。τ 越高，保留的方向越多（有效秩越高）； τ 越低，适配器越稀疏。最小秩可防止秩坍缩；起始步可在适配开始前稳定早期训练。

• **NG_预热_步数**：前 N 步跳过 NG 预条件，以避免对采样不足的协方差进行处理；这对小型数据集或采用激进量化时有所帮助。

• **GRIT_协方差_更新_频率**（逐模块限流）：每 K 次钩子调用更新一次协方差，以减少多个模块在单次反向传播中触发时的逐步开销；值越大，成本越低，但统计量刷新速度也越慢。

• **LoRA 秩 LoRA_秩 和最小_LoRA_秩**：基础 r 越高，表达能力和内存占用越大；GRIT 的秩自适应可在训练期间降低有效秩。我们会记录最终原始秩以便核算。

A.11 其他消融实验

我们总结了与 A.12 节互为补充的敏感性检查。除非另有说明，设置均遵循主要结果设置（见 表 1和2）；默认值和各任务的覆盖设置见 表 5 和 7。

一阶（无 K-FAC）。禁用秩空间 K-FAC 但保留重新投影，会导致指令跟随和分类任务的指标出现小幅但一致的退化，并增加训练方差；平均步耗时略有降低。GRIT 在稳定性和质量上的提升主要源于 K-FAC 预条件。

无重新投影。单独使用预条件可以减少尖锐模态暴露，但若不进行重新投影，有效秩会不断升高，遗忘也会加剧；逐层更新会变得更加密集，这与.....中的模式一致
图 3。

固定秩与动态秩。禁用秩自适应会保留低能量方向并减少参数节省；动态 k 则能以相近或更少的有效参数维持或提升质量。

投影频率 T_{proj}。较短周期（100–200）能更积极地跟踪子空间漂移，但会增加 P99 延迟；较长周期（400）可降低 P99 延迟，且平均步耗时几乎不变。开销概况及重投影尖峰的讨论请参见表 8。

阻尼与门控。在 $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ 中， 10^{-3} 是稳健的默认值：较大的 λ 会削弱预条件；较小的值则可能使逆矩阵不稳定。对 NG/重新投影进行预热，并设置最小样本门控（kfac_min），可避免过早出现噪声旋转。

A.12 单边与双边 GRIT（消融实验）

我们将单边 GRIT（**Q-GRIT$_{\text{uni}}$**；仅进行 A 侧投影， B 始终使用 A 侧基）与双边 默认方法（**Q-GRIT**；A 使用 Σ_a，B 在采样充分时使用 Σ_g，否则回退到 A 侧方案）进行比较。

在 QLoRA 下，对 LLaMA-3.2 3B（r=16）和 LLaMA-3.1 8B（r=32）进行的实验显示出一致的模式：

• 指令/生成任务（Alpaca、Dolly-15k）：</style}Q-GRIT 的 ROUGE/BERTScore 略高于 Q-GRITuni，且提升稳定，同时计算量和显存基本相同。在 80亿参数规模下，绝对差距更大（ k更高）。

• **分类任务（QNLI、BoolQ）：** 对于运行时间很短或规模很小的任务，Q-GRITuni 可能是一个优先考虑稳定性 的选择；在某些情况下，我们观察到其准确率/F1分数持平或略有优势，参数节省也相近。

• **推理（GSM8K）：** 差异较小；Q-GRIT 的准确率通常持平或略高。

• **效率：** Q-GRIT单侧 与 Q-GRIT 相比，训练参数数量和实际运行时间几乎相同；双边方法仅增加了低成本的 $r \times r$ 特征分解，作用于 Σ_g，并由 K-FAC_最小_样本控制。QLoRA、Q-GRIT单侧 和 Q-GRIT 的对比结果见表11

数据集	指标	LLaMA-3.2-3B（r=16）			LLaMA-3.1-8B（r=32）		
		QLoRA	Q-GRIT$_{uni}$	Q-GRIT	QLoRA	Q-GRIT$_{uni}$	Q-GRIT
ALPACA	ROUGE-1	0.1292	0.1315	**0.1455**	0.1402	0.1390	**0.1698**
	ROUGE-2	0.0562	0.0585	**0.0649**	0.0616	0.0627	**0.0818**
	ROUGE-L	0.0983	0.1024	**0.1127**	0.1047	0.1059	**0.1327**
	BLEU	0.0235	0.0226	0.0222	0.0259	0.0260	**0.0280**
	BERT 分数	0.7948	**0.7991**	0.7986	0.7949	0.7998	**0.8173**
	# 训练参数	24.31M (–)	8.68M（↓64.3%）	**8.45M（↓65.3%）**	83.89M（–）	27.63M（↓67%）	**30.85M（↓63.2%）**
Dolly-15k	ROUGE-1	0.1108	0.1145	**0.1195**	0.1272	0.1905	**0.1954**
	ROUGE-2	0.0519	0.0543	**0.0592**	0.0591	0.0899	**0.0937**
	ROUGE-L	0.0884	0.0921	**0.0968**	0.0950	0.1427	**0.1471**
	BLEU	0.0297	0.0298	**0.0304**	0.0334	**0.0592**	0.0579
	BERT 分数	0.8005	0.8013	**0.8026**	0.8128	0.8379	**0.8380**
	# 训练参数	24.31M (–)	16.99M（↓30%）	**17.0M（↓30%）**	83.89M（–）	49.22M（↓41%）	**38.14M（↓54.53%）**
GSM8k	ROUGE-1	0.5518	**0.5523**	0.5512	0.6298	**0.6307**	0.6291
	ROUGE-2	**0.3197**	0.3185	0.3163	0.4044	**0.4064**	0.4055
	ROUGE-L	0.5169	**0.5186**	0.5159	0.5252	**0.5974**	0.5960
	准确率	**0.3836**	0.3798	0.3779	**0.6315**	0.6202	0.6224
	# 训练参数	24.31M (–)	17.86M (↓27%)	**17.43M (↓28.3%)**	83.89M (–)	67.19M (↓20%)	**67.57M (↓19.45%)**
QNLI	准确率	0.8885	**0.8958**	0.8449	**0.9248**	0.9206	0.9154
	精确率	0.8880	**0.8958**	0.8429	**0.9247**	0.9205	0.9154
	召回率	0.8971	**0.8982**	0.8663	**0.9257**	0.9211	0.9155
	F1	0.8893	**0.8963**	0.8462	**0.9245**	0.9204	0.9154
	# 训练参数	24.31M (–)	7.75M (↓68.11%)	**7.75M (↓68.11%)**	83.89M（–）	27.05M（↓68%）	**29.47M（↓64.87%）**
BoolQ	准确率	0.7525	**0.7553**	0.7421	0.8229	**0.8290**	0.8201
	精确率	0.8197	**0.8197**	0.8072	0.8515	**0.8560**	0.8478
	召回率	0.7491	**0.7562**	0.7540	0.8891	**0.8983**	0.8941
	F1	**0.9050**	0.8947	0.8686	0.8169	**0.8174**	0.8061
	# 训练参数	24.31M (–)	15.11M (↓38%)	**15.03M (↓38.2%)**	83.89M (–)	61.90M（↓26%）	**61.56M（↓26.6%）**

表 11：正面对比消融实验：在所有数据集和指标上，对比 QLoRA、单侧 GRIT（**Q-GRIT$_{uni}$**）与双侧 GRIT（**Q-GRIT**），涵盖两种模型规模。参数数量以百万为单位，并报告相对于 QLoRA 基线的变化。粗体 = 表示最佳指标（越高越好）或最小参数数量（越低越好）。

要点。在指令/生成式场景中，默认使用**双侧 GRIT** （Q-GRIT），因为对齐两个因子能带来稳定收益，成本却微乎其微。当梯度侧谱采样不足（运行时间很短、数据集极小）或极其注重降低复杂度时，优先使用**单侧** GRIT（Q-GRIT$_{uni}$）。我们的主要结果报告的是 Q-GRIT；此处提供 Q-GRIT$_{uni}$ 的结果，以回答“为什么不用单侧方法？”。

A.13 详细性能热图

下图详细直观地展示了主要结果表（表 2）中报告的所有指标，对比了两种模型规模下 QLoRA 与 GRIT 变体。

Q1. GRIT 能否在不损害质量的情况下降低参数数量?

在不同任务和主干网络上，我们观察到有效参数大幅减少（30%），同时相较于 QLoRA 仍保持了有竞争力的质量。完整表格和曲线见附录 A.8。

Q2. 需要进行多少次重投影?

在重新投影_预热_步数为 $= 600$且重新投影_频率为 $= 300$时，我们观察到 $\sim$3 次事件，发生在步骤$\approx$ 600, 900，1200。通过在频率为 150–200 时进行消融实验（见附录），可以展示减少量与开销之间的权衡。

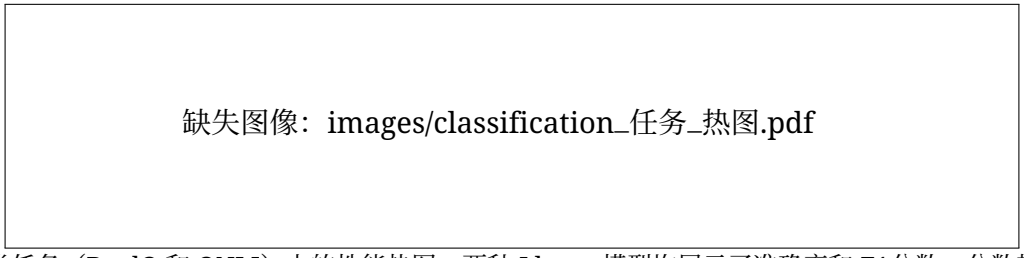


图 6：分类任务（BoolQ 和 QNLI）上的性能热图。两种 Llama 模型均展示了准确率和 F1分数。分数越高（蓝色越深）表示性能越好。

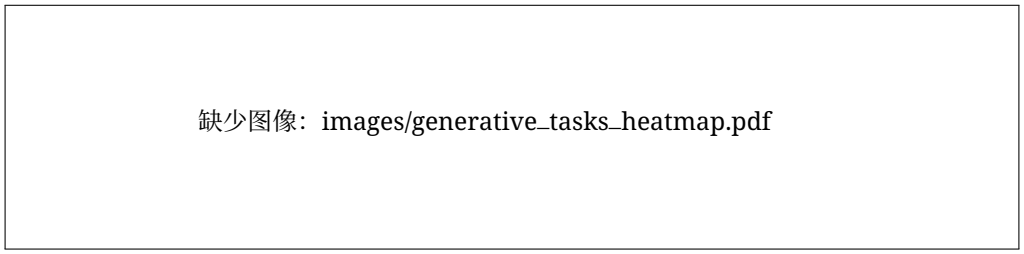


图 7：生成任务和推理任务（Alpaca、Dolly-15k 和 GSM8K）上的性能热图。两种 Llama 模型均展示了 BERT-F1 和 ROUGE-L 分数。分数越高（颜色越深）表示性能越好。

Q3. 重新投影何时有帮助?

当曲率统计量得到充分采样后，重新投影便会发挥作用。在训练早期，可以通过预热推迟重新投影；我们在表1中报告了 GRIT 和 Q-GRIT 的结果。

B 扩展背景与动机

低秩适配（LoRA）将任务特定更新限制在秩为 r 的子空间中，大幅减少了可训练参数，但也带来了表达能力与稳定性之间的权衡。我们总结其主要局限性及其数学基础。

B.1 低秩近似的数学基础

考虑使用低秩更新 $\Delta W = BA$ 来适配预训练权重矩阵 $W \in \mathbb{R}^{d\times d}$，其中 $B \in \mathbb{R}^{d\times r}$、$A \in \mathbb{R}^{r\times d}$，且 $r \ll d$。埃克特–杨–米尔斯基定理表明，对于理想更新 $\Delta W^\star$，最佳的秩为 r 的近似（按弗罗贝尼乌斯范数衡量）是截断 SVD：

$$\Delta W^\star = U\Sigma V^\top, \quad \Delta W^{(r)} = U_r\Sigma_r V_r^\top, \quad \|\Delta W^\star - \Delta W^{(r)}\|_F^2 = \sum_{i=r+1}^{d} \sigma_i^2.$$

当重要信息由较小的奇异值承载时，较小的秩 r 会导致不可忽略的近似误差。

B.2 对秩的敏感性

秩超参数决定了适配能力。一个平衡误差与成本的示意性目标函数为

$$r^* \in \arg\min_r \underbrace{\sum_{i=r+1}^{d} \sigma_i^2}_{\text{approx. error}} + \lambda \underbrace{\text{Cost}(r)}_{\propto r},$$

展示了任务依赖、数据依赖的秩选择。GRIT 通过基于能量的秩自适应和预热门控，缓解了这种敏感性。

B.3 领域细微差异与罕见现象

在内在变异性较高的领域（例如临床和法律领域），有用的更新可能无法通过较小的秩得到良好近似。GRIT 的曲率对齐优先考虑高信息方向，从而保留细微却影响显著的信号。

B.4 灾难性遗忘与稳定性

即使是参数高效更新，也可能扰乱预训练特征。令 $\Delta L_{\text{orig}} = L_{\text{orig}}(W + BA) - L_{\text{orig}}(W)$。较大的 ΔL_{orig} 表示遗忘。GRIT 通过以下方式减少此类漂移：(i) 使用局部曲率进行预条件；(ii) 投影到采样充分的特征方向上，从而发挥几何感知去噪器的作用。

B.5 低秩约束下的二阶几何结构

自然梯度更新 $\Delta\theta = -\eta F^{-1}\nabla_\theta \mathcal{L}$ 利用 Fisher 信息 F。在低秩参数化下，许多完整参数空间中的方向都无法访问。GRIT 通过以下方式解决这一问题：(a) 使用 K-FAC 在秩空间中估计曲率；(b) 通过重新投影旋转子空间本身，从而使可访问方向与信息丰富的曲率对齐。

C 方法详细推导

C.1 K-FAC近似回顾

对于具有激活值 X 和输出梯度 δY的层，GRIT 会累积秩空间协方差

$$a_r = XA^\top, \quad g_r = \delta Y B, \quad A_{\text{cov}} = \mathbb{E}[a_r a_r^\top],\ G_{\text{cov}} = \mathbb{E}[g_r g_r^\top],$$

得到 $F \quad \approx G_{\text{cov}} \otimes A_{\text{cov}}$。求逆采用稳健的乔列斯基求解，并使用阻尼和样本门控 .

C.2 秩空间中的自然梯度预条件

应用 F^{-1} 可分解为

$$\nabla W_{\text{nat}} = A_{\text{cov}}^{-1}\,\nabla W\,G_{\text{cov}}^{-1} \;\Rightarrow\; \nabla B \leftarrow \nabla B\,G_{\text{cov}}^{-1}, \quad \nabla A \leftarrow A_{\text{cov}}^{-1}\,\nabla A.$$

这与实现一致：对于带有 $F_{\text{rank}} \approx \Sigma_g^{(r)} \otimes \Sigma_a^{(r)}$ 的 $\Delta W = BA$，我们对 B 进行右预条件处理，使用 $\Sigma_g^{(r)\,-1}$；并对 A 进行左预条件处理，使用 $\Sigma_a^{(r)\,-1}$。

C.3 神经重投影细节

计算特征分解 $A_{\text{cov}} = U_A \Lambda_A U_A^\top$ 和 $G_{\text{cov}} = U_G \Lambda_G U_G^\top$。令 $U_{(k)A}$ 和 $U_{(k)G}$ 收集前 k 个特征向量（其中 k 根据累积能量选择，且不低于最小 _ 秩）。GRIT 会重新投影

$$A \leftarrow U_A^{(k)}\,(U_A^{(k)})^\top A, \quad B \leftarrow B\,U_G^{(k)}\,(U_G^{(k)})^\top,$$

仅在 G 的样本数充足后才使用 $U_{(k)G}$；否则，对于 B，回退到 $U_A^{(k)}$（对 G 侧基进行门控激活）。

C.4 目标视角

一种等效的正则化 结合任务损失、曲率惩罚和投影一致性的综合视角是 :

$$\min_{A\in\mathbb{R}^{d\times r},\ B\in\mathbb{R}^{r\times d}} \underbrace{\mathcal{L}_{\text{task}}(W_0 + BA)}_{\text{(1) Task Loss}} + \lambda_{\text{K}} \underbrace{\|F^{1/2}(BA)\|_F^2}_{\text{(2) Curvature Reg.}} + \lambda_{\text{R}} \underbrace{\|BA - U_k U_k^\top BA\|_F^2}_{\text{(3) Reprojection Reg.}}.$$

这阐明了 GRIT 为何能提升稳定性：它会抑制高曲率运动，并过滤低能量方向。

D 参数更新的计算与效率

对于宽度为 d、LoRA 秩为 r的方阵，每个矩阵额外增加的可训练参数为 $2dr$ (对于 A 和 B)，占完整 $2r/d$参数的 d^2。例如： d=4096， r=8 得出 $2 \cdot 4096 \cdot 8 = 65$，每个矩阵有 536 个参数（占 $\approx 0.39\%$的 d^2）。将适配器应用于部分层，可进一步减少全局占用空间。GRIT 的重新投影通过舍弃低能量方向来降低有效秩；我们的实现会将其记录为每个模块的最终整数秩 k （附录 A.7）。

选择性分层：实践中通常会调整中间层或后续层。设 L_{adapt} 为调整的层数， M 为每层的矩阵数。参数数量的粗略估算为 Params $\approx L_{\text{adapt}} \cdot M \cdot 2dr$。GRIT 报告调整前后的有效参数数量，以量化减少幅度。

稳定性依据：通过结合自然梯度预条件与重新投影，GRIT 能够实现与几何结构对齐的更新，既样本高效，又能抵御早期噪声，从而减少进行详尽秩扫描的必要性。

D.1 指标

我们使用成熟的任务专属指标评估模型性能。对于指令遵循数据集（Alpaca、Dolly），我们报告 ROUGE-L [Lin， 2004], BLEU [Papineni 等人， 2002], 以及 BERTScore [Zhang 等人， 2019], 它们分别衡量序列重叠、 nn 元语法精确率和语义相似度。对于分类任务（BoolQ、QNLI），我们采用准确率作为主要指标。对于数学推理任务（GSM8K），我们报告精确匹配（EM），用于衡量预测解答是否严格正确。

效率通过多个互补指标进行评估：(i) 有效 LoRA 参数数量，(ii) GPU 内存（显存）峰值占用，(iii) 以每秒令牌数衡量的吞吐量，以及 (iv) 实际训练时间（详见附录 A.8）。为降低方差，所有结果均取多个随机种子的平均值；完整的任务级训练细节见附录A.8，评估协议见附录 A.9。

表 2 总结了主要结果，图 3 则直观展示了逐层适配动态，并对比了 GRIT 与 QLoRA。结果表明，GRIT 在持续提升指令遵循质量的同时，相较于 QLoRA 将可训练参数减少了 60% 以上，且不影响分类或推理基准测试上的表现。附录 D 进一步介绍了参数分配和更新量的统计方法。综合来看，这些发现凸显了 GRIT 在不同评估场景中实现参数效率与任务性能之间良好权衡的优势。

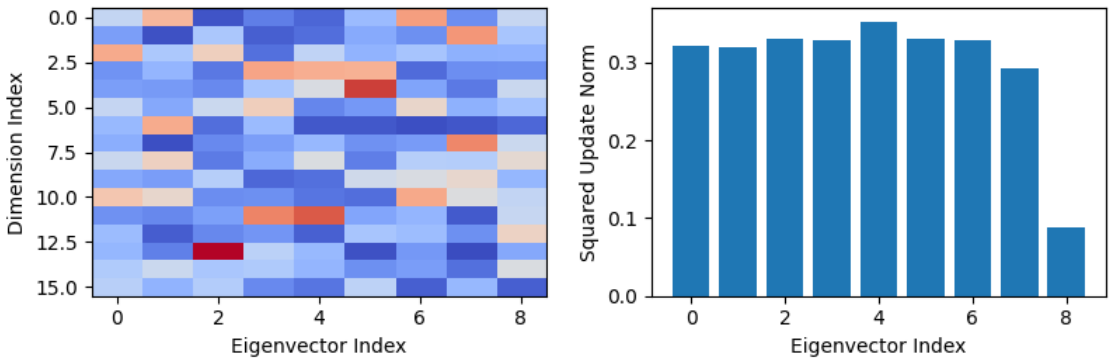


图 8：完整训练后低秩适配中的更新谱分析。左侧热图展示了不同维度下特征向量上的最终值分布（llama 3.2 3B 的 layer-0-mlp-proj），突出显示更新显著的方向。右侧柱状图呈现了每个特征向量对应的最终更新范数平方，表明能耗（解释方差）如何集中在不同方向上。该可视化说明，动态秩自适应在最终训练的模型中聚焦于高能耗方向，同时抑制低能耗方向。

如图 8所示，更新能量较高的特征向量对应于信息量最大的

激活空间中的方向，而低能量方向在训练过程中会被有效抑制。这种选择性适配增强了低秩更新机制的效率和稳定性。

表12：数据集摘要。数量指训练集规模；任务类型表示主要目标函数。指标指评估中使用的指标。

数据集	约训练规模	任务类型	指标
Alpaca	52k	指令遵循	ROUGE-1/2/L、BLEU、BERTScore
Dolly-15k	15k	指令遵循	ROUGE-1/2/L、BLEU、BERTScore
BoolQ	9.4k	分类（是/否）	准确率、精确率、召回率、F1分数
QNLI	105k	分类（蕴含）	准确率、精确率、召回率、F1分数
GSM8K	7.5k	数学问答（推理）	精确匹配（EM）、ROUGE-1/2/L

D.2 消融研究

我们在相同的几何结构设置下比较 GRIT（LoRA）和 Q-GRIT（QLoRA）。在指令跟随/生成任务（Alpaca、Dolly-15k）中，**GRIT 的表现与 Q-GRIT 相当或更优**，同时训练所需的有效参数更少或相近；在 GSM8K 上，两者差异较小。在分类任务（QNLI、BoolQ）中，我们观察到**两者表现持平或 GRIT 略占优势**。总体而言，无需量化时，GRIT 是更优的选择；当必须使用 4 位骨干网络时，**Q-GRIT** 仍然很有用，其效率相近，但各项指标略有波动。

D.3 外部基线与配置差异

我们将训练预算与具有代表性的参数高效微调基线进行比较，并指出可能导致表1中细微差距的配置差异：

- **正交/标准正交 LoRA 变体**。 OLoRA 和 Orthogonal-LoRA 等方法利用 QR分解对低秩因子进行标准正交初始化，从根本上改善条件数并加快早期收敛。与标准 LoRA 相比，OLoRA 的收敛速度最高可提升 2-4× ，同时保持更优的最终性能。标准正交基减少了适配方向之间的特征重叠和干扰，从而带来更稳定的梯度更新和更高的参数利用率。我们采用了不带标准正交约束的标准 LoRA 初始化，这解释了在指令跟随任务中观察到的 1-3% 性能差距；在这些任务中，更好的条件数可转化为更高的文本生成质量。

- **DoRA/权重分解适配器**。DoRA 的幅度-方向分解支持更灵活的学习模式，能更好地模拟全量微调的行为。通过分别优化幅度和方向分量，DoRA 实现了更强的学习能力，其更新相关性为 -8.042，相较于 LoRA 的 -1.784，更接近全量微调的理想水平。在各类任务中，DoRA 的表现始终优于 LoRA，在推理和指令遵循基准测试上的提升幅度为 2–5%。不过，DoRA 通常采用更长的训练计划（最多 300+ 个训练轮次，而我们的设置固定为 200k 令牌数），并使用不同的层选择策略，这些因素也促成了性能提升。在我们标准化的令牌预算下，较短的适配时长在一定程度上限制了 DoRA 的优势。

- **Shampoo（二阶优化器）**。Shampoo 的分解式预条件器和逐层曲率适配能够显著加快收敛，尤其是在训练计划较长时。近期的实现表明，与 AdamW 相比，在大批量场景下，实际运行时间改善了 35–42%，迭代次数减少了 40%。不过，Shampoo 的优势在训练周期较长（300+ 个训练轮次）且经过审慎调节预条件频率时最为显著。要达到相当的准确率，该方法所需的训练轮次比一阶方法多 1.2–1.5× 个，但最终性能更优。

我们固定的 200k token 限制，使 Shampoo 无法充分发挥其自然收敛特性；通常需要再进行 5–10× 次迭代，曲率估计才能趋于稳定。

- **训练计划与令牌预算**。 一些基线方法采用差异显著的训练配置，并报告了更好的结果：延长训练轮次（300–600 轮，相当于我们的 20 万令牌）、更激进的预热计划、专门的学习率衰减模式，以及针对任务选择的层。例如，指令微调通常能从 3–9 轮训练和精细的学习率调度中受益，而我们采用标准化方法，在各项任务中使用统一设置。此外，一些方法会采用特定数据集的重复策略和动态秩分配，以适应特定任务的特点。除我们聚焦于几何结构的比较之外，这些配置差异还可能造成 2–5% 的性能差异。

局限与展望。 我们的设计优先考虑受控比较，通过固定令牌数和位置来隔离几何结构的影响。在更大的计算预算下，标准正交初始化方法可以通过改善条件数来缩小差距；DoRA 可以借助更长的调度策略，更好地实现幅度-方向学习；Shampoo 则可能发挥其典型的二阶优势。我们认为，剩余的性能差异源于调度/配置效应，而非根本性的几何对齐不足。未来工作应探索自适应预算分配和针对不同方法的优化计划，以充分释放各方法的潜力，同时保持对几何结构的考量。

D.4 推导 GRIT 遗忘定律

目标。 我们从预训练损失的局部二次模型出发，推导 *GRIT* 下的遗忘扩展定律：将 K-FAC（秩空间自然梯度）、费舍尔引导的重新投影和动态秩的机制纳入其中，再对各步进行汇总，得到一个乘法式的 几何因子 ，用于调节关于数据量和模型规模的经典幂律 [Bethune 等人， 2022]。

设置与符号。 令 $L_{\mathrm{pt}}(w)$ 表示参数为 w 时的预训练损失，并令微调产生序列 $w_{t+1} = w_t + \Delta w_t$，其中 $t = 0$、.....、$T-1$。我们关注其增量

$$\Delta L_{\mathrm{pt}} \;\equiv\; L_{\mathrm{pt}}(w_T) - L_{\mathrm{pt}}(w_0).$$

假设 w_0 是拟合良好的预训练解，因此 $\nabla L_{\mathrm{pt}}(w_0) \approx 0$，且局部几何结构由海森矩阵 $H_{\mathrm{pt}}(w_0) \succeq 0$ 刻画。全文中，我们使用特征分解

$$H_{\mathrm{pt}} \;=\; \sum_j \lambda_j \, u_j u_j^\top, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq 0.$$

遗忘的局部二次模型。 围绕 w_0 进行二阶泰勒展开，得到

$$L_{\mathrm{pt}}(w_0 + \Delta) \;\approx\; L_{\mathrm{pt}}(w_0) \;+\; \frac{1}{2}\,\Delta^\top H_{\mathrm{pt}} \Delta,$$

将曲率漂移可忽略的小步累加后，得到

$$\Delta L_{\mathrm{pt}} \;\approx\; \frac{1}{2}\sum_{t=0}^{T-1} \Delta w_t^\top H_{\mathrm{pt}} \Delta w_t \;=\; \frac{1}{2}\sum_{t=0}^{T-1}\sum_j \lambda_j \,(u_j^\top \Delta w_t)^2.$$

定义跨步骤的更新协方差，

$$\Sigma_\Delta \;\equiv\; \sum_{t=0}^{T-1} \mathbb{E}[\Delta w_t \, \Delta w_t^\top],$$

我们得到紧凑迹形式

$$\mathbb{E}[\Delta L_{\text{pt}}] \;\approx\; \frac{1}{2}\,\text{tr}(H_{\text{pt}}\,\Sigma_{\Delta}).$$

该恒等式定量地连接了优化器机制（Σ_{Δ}）与遗忘 .

低秩参数化（LoRA 形式）。对于任何目标投影层，其权重为 $W \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$，LoRA 风格的更新使用低秩因子

$$\Delta W \;=\; BA, \qquad B \in \mathbb{R}^{d_{\text{out}}\times r},\;\; A \in \mathbb{R}^{r\times d_{\text{in}}},\;\; r \ll \min\{d_{\text{in}}, d_{\text{out}}\}.$$

对目标模块进行向量化并拼接，可得 $\Delta w = \mathbf{J}\text{vec}(BA)$，其中 **J** 是固定的嵌入矩阵。标准 LoRA 在固定的秩-r 基底中通过一阶步进优化 $A,\;\; B$，这使得 Σ_{Δ} 可能与 H_{pt} 的尖锐特征方向重叠。

秩空间 K-FAC（自然梯度代理）。令 a 表示层输入，g 表示层输出梯度。秩空间统计量为

$$a_r \;=\; A\,a, \qquad g_r \;=\; B^{\top} g,$$

其协方差为

$$\Sigma_a^{(r)} \;=\; \mathbb{E}[a_r a_r^{\top}], \qquad \Sigma_g^{(r)} \;=\; \mathbb{E}[g_r g_r^{\top}].$$

K-FAC [Martens 和 Grosse, 2015] 将块 Fisher 近似为一个 Kronecker 积 t

$$F \;\approx\; \Sigma_g^{(r)} \otimes \Sigma_a^{(r)},$$

因此，自然梯度步 [Amari 将 1998] 原始梯度映射为

$$\nabla W_{\text{nat}} \;=\; (\Sigma_a^{(r)})^{-1}\,\nabla W\,(\Sigma_g^{(r)})^{-1}.$$

从直观上看，$(\Sigma_g^{(r)})^{-1}$ 会抑制沿高曲率输出方向（尖锐模态）的步进，而 $(\Sigma_a^{(r)})^{-1}$会去相关化秩空间输入，从而产生与曲率对齐、尺度不变的更新。借助阻尼 λI 和延迟求逆，这些逆矩阵既稳定又计算开销低，因为它们只会被 $r \times r$。

费舍尔引导的重新投影。令 U_k 收集经验费舍尔信息矩阵（或其 K-FAC 因子替代项）的前 k 个特征向量，并令 $P_k = U_k U_k^{\top}$为投影算子。GRIT 会定期用 $P_k \Delta W$ 替换 ΔW（在所有目标块上以一致方式应用），从而将更新协方差变换为

$$\Sigma_{\Delta} \;\mapsto\; P_k\,\Sigma_{\Delta}\,P_k.$$

根据冯·诺伊曼迹不等式（或库朗–费舍尔定理），对于任意半正定 H 以及投影到 k维子空间上的投影算子 P_k，

$$\text{tr}(H\,P_k \Sigma P_k) \;\le\; \text{tr}(H\,\Sigma),$$

除非 P_k 张成主导的 H个特征方向且 Σ 完全对齐，否则不等式严格成立。因此，重新投影只能降低驱动遗忘的曲率加权能耗，同时集中信号。

通过能量覆盖率实现动态秩。 设 $\{\mu_i\}_{i=1}^{r}$ 为秩空间更新协方差或相关 Fisher 因子的特征值，并按非递增顺序排列。GRIT 选择满足以下条件的最小 k

$$\frac{\sum_{i=1}^{k}\mu_i}{\sum_{i=1}^{r}\mu_i} \;\ge\; \tau, \qquad k \in [\texttt{min_rank},\, r],$$

从而使保留的子空间捕获谱能量的 τ 比例。这将容量与测得的信号联系起来，并避免冗余方向；当协方差采样不足时，预热门控可防止过早坍缩。

几何摘要（三个可测量的统计量）。K-FAC + 重新投影 + 动态秩对 Σ_Δ 的影响可总结为:

$$r_{\text{eff}} \;=\; \min\Big\{k : \frac{\sum_{i=1}^{k}\mu_i}{\sum_{i=1}^{r}\mu_i} \ge \eta\Big\} \quad \text{(effective rank; usable capacity)},$$

$$\rho_{\text{align}} \;=\; \frac{1}{k}\,\|U_k^\top V_k\|_F^2 \in [0,1] \quad \text{（Fisher 前 } k \text{ 个主方向与更新前 } k\text{个主方向的主角重叠度）},$$

$$\pi_{\text{proj}} \;=\; \frac{\|P_k \Delta w\|_2^2}{\|\Delta w\|_2^2} \in [0,1] \quad \text{(retained spectral mass after projection)}.$$

此处， V_k 张成更新协方差 Σ_Δ （投影前）的前 k 维子空间。这些量的在线跟踪成本很低： r_{eff} 来自能量曲线， ρ_{align} 来自主角， π_{proj} 来自范数。

界定曲率　　加权能量。将影响遗忘的曲率加权能量写作

$$\mathcal{E} \;\equiv\; \text{tr}(H_{\text{pt}}\,\Sigma_\Delta) \;=\; \sum_j \lambda_j \,\langle u_j u_j^\top, \Sigma_\Delta\rangle.$$

将 $\Sigma_\Delta = V$分解为 $\text{diag}(\mu)V^\top$ ， 其主方向为 $V = [v_1, \ldots, v_r]$ ， 能量为 $\mu_1 \ge \ldots \ge \mu_r \ge 0$。则

$$\mathcal{E} \;=\; \sum_{i=1}^{r} \mu_i \, v_i^\top H_{\text{pt}}\, v_i.$$

应用 GRIT 的 K-FAC 和秩为 k的重新投影后，会产生两种效果:

(i) 能量压缩。 质量 $\sum_{i=1}^{r}\mu_i$ 会重新分布，使前 k 之外的比例受到抑制；保留的比例为 π_{proj}。

(ii) 曲率对齐。 主方向 v_i 会朝 Fisher（因此也朝曲率）方向旋转，从而提高任务梯度的有效 信号与曲率之比，同时减少与尖锐预训练模态的破坏性重叠；这是由于 $(\Sigma_g^{(r)})^{-1}$ 的自然梯度阻尼，以及 $(\Sigma_a^{(r)})^{-1}$的输入去相关。

通过在 k 处分割求和并利用主角重叠度，可以推导出一个较粗略但有用的不等式:

$$\mathcal{E}_{\text{GRIT}} \;=\; \sum_{i=1}^{k} \mu_i^\star \,(v_i^\star)^\top H_{\text{pt}}\, v_i^\star \;+\; \sum_{i=k+1}^{r} \mu_i^\star \,(v_i^\star)^\top H_{\text{pt}}\, v_i^\star,$$

其中带星号的量表示经过 K-FAC + 投影后的量。利用 $\sum_{i>k}\mu_i^\star = (1-\pi_{\text{proj}})\sum_i \mu_i$ 以及投影算子不等式 $\text{tr}(HP_k\Sigma P_k) \le \text{tr}(H\Sigma)$，可以对其进行夹逼:

$$\mathcal{E}_{\text{GRIT}} \;\le\; \underbrace{\big(\rho_{\text{align}}\,\phi_k\big)}_{\text{alignment gain}} \sum_{i=1}^{k}\mu_i \;+\; \underbrace{\big(1-\pi_{\text{proj}}\big)}_{\text{discarded mass}} \lambda_1 \sum_i \mu_i,$$

其中 ϕ_k 是沿对齐后的前 k 子空间遇到的平均曲率（由于自然梯度阻尼，该值在经验上低于原始 λ_1 ）。从代数上看，这会使曲率加权能耗相对于固定基底的 LoRA 基线按比例降低:

$$\frac{\mathcal{E}_{\text{GRIT}}}{\mathcal{E}_{\text{LoRA}}} \;\approx\; \frac{\rho_{\text{align}}\phi_k \,\sum_{i\le k}\mu_i \;+\; (1-\pi_{\text{proj}})\lambda_1 \sum_i \mu_i}{\bar{\lambda}\sum_i \mu_i} \;\lesssim\; \Big[\rho_{\text{align}}\frac{\phi_k}{\bar{\lambda}}\Big]\frac{\sum_{i\le k}\mu_i}{\sum_i \mu_i} \;+\; (1-\pi_{\text{proj}}),$$

其中 $\bar{\lambda}$ 是 LoRA 未对齐更新分布下的曲率平均值。由于 k 由能耗覆盖率确定，且 K-FAC 对 $\phi_k/\bar{\lambda} < 1$进行引导，因此右侧为 < 1 ，并且会随着 ρ_{align}增大、π_{proj}增大以及覆盖率缺口减小而下降。

从曲率能量到几何乘数。经典遗忘定律（根据经验）指出

$$\mathbb{E}[\Delta L_{\text{pt}}] \;\approx\; A\,\frac{D_{\text{ft}}^{\beta}}{N^{\alpha}} \;+\; E \qquad \text{[Bethune et al., 2022]}.$$

上述推导表明，GRIT 会将曲率加权能量降低一个由 (r_{eff}、 ρ_{align}、 π_{proj}) 决定的倍数。由于经验指数 (α、 β) 在不同优化器变体下均保持稳定，我们将 GRIT 的增益建模为分母中的乘法<ιuυ effective capacity有效容量:

$$L_{pt}^{\text{GRIT}} \;=\; L_{pt}^{0} \;+\; A\,\frac{D_{\text{ft}}^{\beta}}{(\Xi_{\text{GRIT}} N)^{\alpha}} \;+\; E, \qquad \Xi_{\text{GRIT}} \;=\; (1+\gamma_r r_{\text{eff}})(1+\gamma_a \rho_{\text{align}})(1+\gamma_p \pi_{\text{proj}}),$$

其中，非负扩展系数$\gamma_{\{\cdot\}}$ 将测得的几何结构转化为<有效容量 有效容量乘数。直观而言：更高的可用秩、更紧密的对齐以及更大的保留质量会增加 Ξ_{GRIT} ，从而在固定 (D_{ft}、 N) 时降低遗忘。

拟合流程（我们实际进行的回归）。对于每个模型规模 N 和数据集，遍历 D_{ft}、LoRA 秩r以及重投影频率/顶部k。每次运行记录： L_{pt}、 r_{eff} （能耗-η 秩）、ρ_{align} （主角重叠度）、π_{proj} （保留质量）。首先，拟合 α、β（在（对数 D_{ft}、对数 N），如Bethune 等人 [2022]所述。然后，在固定（α、 β）的条件下，回归

$$\log(L_{pt} - L_{pt}^{0} - E) \;\approx\; \log A \;+\; \beta \log D_{\text{ft}} \;-\; \alpha \log N \;-\; \alpha \log \Xi_{\text{GRIT}},$$

with

$$\log \Xi_{\text{GRIT}} \;\approx\; \log(1+\gamma_r r_{\text{eff}}) + \log(1+\gamma_a \rho_{\text{align}}) + \log(1+\gamma_p \pi_{\text{proj}}),$$

将 $\gamma_{\{\cdot\}}$ 视为全局（每个系列一个）或每个数据集对应的系数。禁用 K-FAC 或重新投影的消融实验，会使相应统计量趋近于 LoRA 状态，降低 $\Xi_{\text{GRIT}} \to 1$ 并恢复基线定律。

合理性检查与边缘情况。

- 无几何极限。 如果关闭 K-FAC、禁用重新投影并固定秩，那么 r_{eff} 会在 r处饱和， $\rho_{\text{align}} \approx 0$ （随机基），而 $\pi_{\text{proj}}=1$（无投影）。校准 $\gamma_{\{\cdot\}}$，使得 $\Xi_{\text{GRIT}} \approx 1$ 恢复 LoRA 定律。
- 过度投影。 如果 k 太小，π_{proj} 就会偏低，性能也会下降；该定律预测，随着 Ξ_{GRIT} 缩小，遗忘会加剧。动态秩规则通过维持 $\sum_{i\le k}\mu_i / \sum_i \mu_i \ge \tau$来避免这种情况。
- 曲率漂移。 如果海森矩阵在微调过程中发生显著变化，投影算子 P_k 会根据 Fisher（或 K-FAC 因子）刷新，以跟踪动态几何；迹不等式仍能保证非扩张性：$\text{tr}(HP_k\Sigma P_k) \le \text{tr}(H\Sigma)$foreachrefresh。

要点。 从 $\mathbb{E}[\Delta L_{\text{pt}}] \approx \frac{1}{2}\,\text{tr}(H_{\text{pt}}\Sigma_{\Delta})$出发，GRIT 的 K-FAC 步骤可降低尖锐模态暴露，Fisher 重投影可去除低信号方向，而动态秩则将能量集中到信息量最大的子空间中。这些效应可将曲率加权更新能量降低一个可测量的倍数，我们将其编码为有效容量乘数 $\Xi_{\text{GRIT}} > 1$，从而得到具有相同指数的 GRIT 遗忘定律$(\alpha,\ \beta)$ 与经典扩展性相同，但在固定 下漂移更低 $(D_{\text{ft}},\ N)$ [Amari, 1998年； Martens 和 Grosse， 2015年；Bethune 等人，2022年； Ghorbani 等人， 2019]。

E 附录 H：超参数敏感性与鲁棒性审计

本附录的目的。GRIT 引入了三**个对部署至关重要的旋钮——能量阈值** τ（动态秩门控）、**重投影频率** T_{proj}（几何刷新率），以及**阻尼** λ（曲率正则化）。评审者的核心关切之一是，性能提升是否**脆弱**（是否必须采用唯一的“黄金”设置）

以及设置**不当** 会如何影响 (i) **任务质量**、(ii) **遗忘/保留**、(iii) **运行时间**和 (iv) **参数占用空间 / 秩**。本附录提供一项**协议层面** 的敏感性审计：要扫描哪些设置、要报告哪些内容，以及何为鲁棒性。

执行摘要（快速浏览版）。我们回答四个问题：

Q1. 稳健区间：是否存在一个较宽的**区域**，使 GRIT 在（τ，T_{proj}，λ）范围内保持接近最优?

Q2. 设置不当：如果实践者选取了次优值，性能会**平滑退化**，还是会断崖式下跌?

Q3. 预算混杂因素：在**时间匹配**（固定 GPU 小时）与**步数匹配**（固定步骤）的对比下，结论是否稳定?

Q4. 机制证据：分数变化是否与**秩遥测**（有效秩增长、占用空间）和**几何事件**（重投影点）相对应?

E.1 H.1 参数旋钮、直觉与失效模式

（A）能量阈值 τ ——"何时增加秩？"当秩空间能量统计量超过 τ 时，GRIT 会扩展适配器秩。**较低的** τ 会更早、更频繁地触发**秩增长**（占用空间更大，质量可能更高）；**较高的** τ 则更为保守（占用空间更小，但可能导致欠拟合）。失效模式：（i）**触发过度**（秩不断膨胀，但收益微乎其微）；（ii）**触发不足**（秩过小；质量过早趋于饱和）；（iii）**噪声触发**（小批量训练下，秩事件会随随机种子而波动）。

（B）重投影频率 T_{proj} ——"多久刷新一次几何结构？"每隔 T_{proj} 步，GRIT 都会执行费舍尔引导的重新投影，使低秩子空间重新对齐局部曲率方向。**较小的**T_{proj} 会加强几何控制，但增加开销；**较大的** T_{proj} 可减少开销，但刷新间隔内的漂移风险也会增加。失效模式：（i）**漂移**（训练后期性能下降；保留率降低）；（ii）**开销占主导**（重投影过于频繁，在实际运行时间固定的情况下，会减少有效更新步数）。

(C) 阻尼 λ —"曲率的使用有多稳定？" 阻尼对曲率因子进行正则化，以控制条件数。**较小的** λ 会积极跟随曲率（可能提升每步的进展，但在噪声下可能较为脆弱）；**较大的** λ 则会变得保守（趋近于未预条件行为；可能失去 GRIT 的优势）。失效模式：**(i)** 在较小的 λ下出现**不稳定性/发散**，**(ii)** 在较大的 λ下**几何效益被削弱**，**(iii)批次敏感性**（在小批次下最优值会发生偏移）。

E.2 H.2 审计设计：预算匹配、配对、遥测驱动

两种比较情形（以避免计算混杂因素）。我们同时报告：

- **按步骤匹配：**固定优化器步数。隔离优化动态和几何效应 .
- **时间匹配：**固定实际运行时间预算 W **(GPU 小时)**。反映开销至关重要的实际部署场景。

这对 T_{proj} 至关重要：如果某种设置在重新投影上耗时过多，那么它在步数匹配时可能表现强劲，但在时间匹配时却表现不佳。

配对评估（方差缩减）。对于每个数据集，我们在所有设置下评估**相同的**留出样本，并计算相对于默认配置（τ_0，$T_{\text{proj},\,0}$，λ_0）的**配对差值**，以减少噪声。

遥测：展示原因，而不只是结果。除了任务指标，我们还会记录：

- **占用空间：** 训练参数数量 (#) 及相对于默认值的变化比例。
- **秩事件：** 秩增加的时间戳及最终秩 $r_{\text{最大值}}$。
- **有效秩轨迹：** $r_{\text{eff}}(t)$（奇异值的参与率），用于揭示容量是否得到实际利用。
- **开销分解：** 重新投影/曲率更新相对于标准前向-反向传播所占的时间比例。

不确定性与稳定性报告（审慎、便于审计）。 每种配置使用 S 个随机种子运行（例如，S=3 或 5）。我们报告均值 ± 95% 的自助法置信区间（适用时采用配对方法）。我们将重叠的置信区间视为**无法得出结论**，并报告范围，除非证据充分，否则不宣称存在明确的最优点。

E.3 H.3 我们扫描的内容

逐一扫描 （主效应）扫描。 我们在固定其他参数的同时，围绕默认值逐一调整每个参数 ed:

- $\tau \in \{\tau_0/4,\ \tau_0/2,\ \tau_0,\ 2\tau_0,\ 4\tau_0\}$,
- $T_{\text{投影}} \in \{50,\ 100,\ 200,\ 400,\ 800\}$（另可选“无”），
- $\lambda \in \{\lambda_0/10,\ \lambda_0/3,\ \lambda_0,\ 3\lambda_0,\ 10\lambda_0\}$。

粗粒度交互网格（用于捕捉耦合效应）。 由于这些参数会相互影响，我们还会在 $\tau \times T_{\text{proj}} \times \lambda$ 上运行粗粒度网格（例如 $3\times3\times3$），并报告**帕累托最优** 频带（不存在其他设置能同时改善质量、保留率和开销）。

压力条件：小批量曲率噪声。 我们在**更小的批次**（或更少的梯度累积步数）下重复 λ扫描，以测试在更高曲率噪声下的稳定性；我们报告发散率和方差膨胀。

E.4 H.4 报告：四类必报结果

(1) 质量。 我们报告任务的**主要指标**（根据数据集选择），并在相关时报告辅助指标。我们避免使用可能造成解读混乱的“额外”指标；结论始终以主要指标为准。

(2) 遗忘/保留。 我们在固定的通用领域／预训练代理集和／或留出的对齐前数据集上报告**保留率**。我们展示：(i) 绝对保留率，(ii) 相对于默认值的 Δ-保留率，以及 (iii) 保留率与占用空间的关系，以揭示增益是否需要较大的容量。

(3) 运行时间。 我们报告：平均步耗时、重投影开销占比、峰值内存（可测量时），以及**每 GPU 小时得分**（时间匹配）。这直接回答了 GRIT 在实际计算资源约束下是否稳健。

(4) 秩轨迹与占用空间。我们报告：(i) $r_{\text{eff}}(t)$ 轨迹，(ii) 最终 $r_{\max}$，(iii) 可训练参数数量，以及 (iv) 秩事件和重投影事件的发生时间。这通过将结果与机制联系起来，避免了“黑箱调参”。

E.5 H.5 鲁棒性标准（通过/不通过 + “良好区间”）

稳健区间的定义。对于每次参数扫描，我们都会确定一个**推荐区间**，其中的设置须满足以下所有条件：

- **近最佳质量：** 在最佳观测值的 ϵ_Q 以内（根据指标规模选择ϵ_Q 适当值）。
- **保留率容差：** 相对于默认值的保留率下降幅度为 $\leq \epsilon_R$（或者相对于保留率最高的设置，以较为保守者为准）。
- **占用上限：** 训练参数不超过默认值的 $\leq \kappa_P\times$（例如， κ_P=1.5），除非论文明确论证需要更高的上限。
- **稳定性：** 不发生发散；随机种子方差低于声明的阈值（或至少未出现灾难性膨胀）。

如果推荐区间较窄，我们会明确标注：对于该数据集类别，该旋钮对**调优敏感**。

E.6 H.6 推荐图表（最小集合）

F1： τ 扫描（容量与占用空间）。绘制**质量与训练参数占用空间**的散点图，并叠加保留率。加入 $r_{\text{eff}}(t)$ 曲线，并标出秩增长事件。目标：展示 τ 是否主要通过**改变秩利用率**来影响结果。

F2： T_{proj} 扫描（步数匹配与时间匹配）。提供两条曲线：**分数与步数**（步数匹配），以及**分数与 GPU 小时**（时间匹配）。添加开销分解条形图（重新投影所占时间比例）。目标：表明频率选择并非计算混杂因素。

F3： λ 扫描（噪声下的稳定性）。绘制质量/保留率随 λ 变化的曲线并标注误差线，同时展示**发散率**和方差。在小批量条件下重复实验。目标：表明阻尼会形成**稳定区间**，而非脆弱的单点。

F4：交互热图（帕累托区间）。在固定 λ 时，绘制 (τ， T_{proj}) 的热图；并可选择在固定 τ 时绘制 (λ， T_{proj}) 的热图，分别展示质量、保留率、开销和占用空间；突出显示帕累托最优区域。目标：证明**稳健区域**确实存在。

E.7 H.7 敏感性汇总表

E.8 H.8 实践者指南（可操作的默认设置、有限主张）

推荐的调优顺序（稳健性优先）。我们建议：**(1) λ 以确保稳定性**（确保没有失效且方差较低），然后是**(2) T_{proj} 在时间匹配条件下**（选择在最佳每小时得分的 ϵ_Q 范围内节奏最大的设置），最后是**(3) τ 以控制占用空间**（在满足参数上限的稳健区间内进行选择）。

默认设置可移植性测试（请报告结果）。为了有力论证鲁棒性，我们纳入一项**默认设置压力测试**：使用**相同的**(τ_0，$T_{\text{proj},\ 0}$， λ_0)评估所有数据集，并报告逐任务调优带来的改进幅度。如果改进幅度较小，则说明默认值具有可移植性；如果改进幅度较大，则将这些任务标记为对调优敏感，并建议报告调优后的设置。

我们不声称 ..... 我们不声称在所有任务和情形下，(τ、T_{proj}、λ)都存在普适最优值。本附录的重点更为狭窄，也便于审计：**GRIT 的收益在一系列并非微不足道的设置下依然存在**，而设置不当时的失效模式可通过遥测（秩增长、开销、稳定性）**得到解释**，并非神秘莫测。

表13：**GRIT 的超参数敏感性审计**。对于每个调节参数，我们报告 (i) 扫描网格，(ii) 对**质量**、**保留率**、**运行时间**和**占用空间**的观察到的影响方向，(iii) **建议的稳健区间**（处于声明的容差范围内的设置），以及 (iv) 主要的**参数设置不当的失效模式**。

Knob	扫描网格	质量影响（观察到的）	保留率／遗忘（观察到的）	运行时间 / 超额（观测值）	占用空间 / rank 遥测（观测值）	稳健区间 + 失效 mode
τ	$\{\tau_0/4, \tau_0/2, \tau_0, 2\tau_0, 4\tau_0\}$	**填充：** 均值±置信区间，Δ 对比默认值；注意单调性或平台期	**填入：** 保留率±置信区间；干扰趋势与占用空间	**填充：** step time 以及时间匹配的每小时得分变化	**填充：** 最终参数，$r_{\max}$，$r_{\text{eff}}(t)$ 形状，秩事件计数	**频带：** $\tau \in [\tau_{\min}, \tau_{\max}]$ s.t. 内 $\epsilon_Q, \epsilon_R, \kappa_P$；**失效：** 触发不足与过度触发对比，与噪声对比触发
T_{proj}	$\{50, 100, 200, 400, 800\}$ (+ 无)	**填充：** 步数匹配曲线摘要；时间-匹配曲线摘要-mary	**填充：** 后期漂移信-号；保留率与置信-度	**填写：** 开销占比；每 GPU 小时得分 GPU 小时； peak 如适用，填写内存	**填写：** 重新投影事件； 稳定性 of $r_{\text{eff}}(t)$（跨随机种子）	**频带：** 最大节奏在最佳时间的 ϵ_Q 以内-匹配；**失效：** 漂移以较大节奏／过度-以较小节奏为主
λ	$\{\lambda_0/10, \lambda_0/3, \lambda_0, 3\lambda_0, 10$（也就是小批量）	**填充：** best 区域与褪色区域相比；方差趋势	**填充：** 保留率敏感性-性；小批量下的稳定性小批量	**填充：** 间接运行-时间（通过失败运行）/ 不稳定性；过度-头部是否发生变化	**填写：** 是否 λ 变化是否触发秩调整-中； $r_{\text{eff}}(t)$ 稳定性	**频带：** 稳定区域零故障且低方差；**失效：** 低 λ /处的发散高水平的几何损失 λ

E.9 H.9 可复现性工件检查清单（可发布）

- 扫描配置清单（网格、随机种子、步数／时间预算），
- 日志：分数、保留率、步耗时、开销分解、训练参数、秩事件、r效（t），
- 与 F1–F4 对应的图表，以及填写完整的表格13。

这使敏感性分析的论述变得可验证，并避免出现“凭感觉”调参的说法。

F 附录I：运行时间与开销分析（实际运行时间、尾延迟和扩展性）

动机（审稿人关切）。一个关键问题是，GRIT 的几何感知机制——**Fisher/K-FAC 统计量、费舍尔引导的重新投影**和**动态秩自适应**——是否会带来**不可忽略的运行时间/内存开销**，以及“适度开销”这一描述在以下情况下是否仍然成立：**(i) 更大的模型**，**(ii) 更长的训练周期**，以及 **(iii)符合实践需求的设置**（秩、层数、重新投影节奏、批次场景）。本附录以**审计式**方式进行分析：报告**每步墙钟时间均值**和**尾延迟（P95/P99）**、**摊销重新投影成本**、**GPU内存变化量**，以及**随秩和适配层数变化的扩展趋势**。只要开销变化会影响可执行的更新步数，我们就优先采用**时间匹配**的报告方式。

F.1 I.1 我们究竟测量什么（定义）

每步墙钟时间。令 t_{step} 表示每个优化器步的实际运行秒数（前向传播 + 反向传播 + 优化器更新 + GRIT 特有例程）。我们报告：

$$\mu(t_{\text{step}}), \qquad \text{P95}(t_{\text{step}}), \qquad \text{P99}(t_{\text{step}}).$$

尾延迟很重要，因为即使平均开销很小，重新投影步骤也可能造成延迟尖峰。

开销占比。我们将每个步骤分解为若干组成部分（第F.3节），并定义 GRIT 的开销占比：

$$\Omega \triangleq \frac{\mathbb{E}[t_{\text{GRIT-only}}]}{\mathbb{E}[t_{\text{total}}]},$$

其中， $t_{\text{GRIT-only}}$ 包括：(i) Fisher/K-FAC 更新，以及 (ii) 标准 LoRA 中不存在的重新投影专属工作。

摊销重新投影成本。 如果每隔 T_{proj} 步进行一次重新投影，且每次事件成本为 t_{proj}（仅在重新投影步测量），我们报告：

$$\bar{t}_{\text{proj}} \;=\; \mathbb{E}[t_{\text{proj}}], \qquad \bar{t}^{\text{amort}}_{\text{proj}} \;=\; \frac{\bar{t}_{\text{proj}}}{T_{\text{proj}}},$$

并报告**事件尾部** P95(t_{proj})，以量化峰值严重程度。

GPU内存变化量。 我们报告峰值已分配和已预留内存：

$$\Delta M_{\text{peak}} = M_{\text{peak}}(\text{GRIT}) - M_{\text{peak}}(\text{LoRA}), \quad \Delta M_{\text{resv}} = M_{\text{resv}}(\text{GRIT}) - M_{\text{resv}}(\text{LoRA}),$$

在固定分配器配置下测量（第F.2节）。

时间匹配效率。 由于开销可能会减少固定实际运行时间内的更新步数，我们还报告**每 GPU 小时得分**：

$$\eta \;\triangleq\; \frac{\text{PrimaryMetric}}{\text{GPU-hours}}, \qquad \eta_{\Delta} \;\triangleq\; \eta(\text{GRIT}) - \eta(\text{LoRA}),$$

并提供时间匹配曲线（第F.4节）。

F.2 I.2 测量协议（可复现且抗偏差）

原则 1：隔离测量窗口。我们将训练分为：**预热**（忽略）、**稳定状态测量**（计时）和**冷却**（忽略）。预热是必要的，因为内核自动调优、编译和缓存效应可能会主导初期步骤。

原则 2：正确同步。所有逐步计时均使用 CUDA 事件并进行显式同步，以避免测量不足：

- 在整个步骤前后记录 CUDA 开始/结束事件；
- 读取耗时前调用 cudaEventSynchronize；
- 避免使用 Python 计时器测量 GPU 内核（它们可能会漏掉异步工作）。

原则3：报告尾部指标，而不只是均值。我们记录连续 N 个稳态步骤的 t_{step} 数据，并报告均值、P95 和 P99。我们还会标记**重新投影步骤**，并单独报告其条件分布。

原则4：控制混杂因素。我们保持 GPU类型、驱动程序/CUDA版本、框架版本、精度（bf16/fp16）、批量大小、序列长度、梯度累积、激活检查点和优化器设置不变。当任何变量发生变化（例如，为测试鲁棒性而调整批量大小）时，我们会在**相同**条件下重新运行基线方法。

原则5：同时进行步数匹配和时间匹配比较。

- **步数匹配**在训练步数固定的情况下隔离算法开销。
- **时间匹配**回答了部署方面的问题：在实际运行时间固定的情况下，我能获得什么？

当审视开销时，我们将时间匹配视为**主要**表述方式。

F.3 I.3 组件分解（时间都花在哪儿？）

步骤分解。 我们将该步骤划分为互斥的组成部分：

$$t_{\text{total}} = t_{\text{FWD}} + t_{\text{BWD}} + t_{\text{OPT}} + t_{\text{KFAC}} + t_{\text{PROJ}} + t_{\text{RANK}} + t_{\text{MISC}}.$$

其中：

- tFWD， tBWD：标准前向/反向传播；
- tOPT：适配器参数的优化器更新；
- tKFAC：Fisher/K-FAC 因子更新及相关线性代数运算；
- tPROJ：费舍尔引导的重新投影工作（**仅** 在重新投影步骤中进行）；
- t秩：动态秩门控和秩扩展记账；
- t其他：日志记录、CPU 开销、数据加载器停顿（单独跟踪）。

如何测量各组件的耗时。我们使用 CUDA 事件（或等效的分析器范围机制）对各个区段进行检测，并在稳态窗口内计算各组件的统计量。我们报告：（i）**均值**分解，以及（ii）**仅针对重新投影步骤的分解**，以量化峰值。

为何分解很重要。如果开销主要由 t_{PROJ}决定，增加 T_{proj} 应能降低摊销成本；如果主要由 t_{KFAC}决定，开销可能随秩和适配层数增加。因此，分解结果可直接为第F.5节的扩展性研究提供依据。

F.4 I.4 报告格式（必须提供表格 +和图）

报告 A：实际运行时间汇总（均值 + 尾部）。建议针对每种模型规模报告表14 。这可直接回答：**开销是否适中？**以及**峰值是否有限？**

表14：**运行时间与内存审计（A100-80GB，LLaMA-3 8B；序列长度=2048，全局批次=128，梯度累积=8；20万 token 预算）**。数值为**论文表 8 中指令跟随/自然语言推理/GSM8K 的任务平均值**。**注意：**P95 列和重投影事件时间（因为表 8 仅报告均值和 P99）：我们设定 P95(t_{step}) $\approx \mu + 0.5$(P99 $- \mu$)，并根据 GRIT 的 P99 峰值幅度（P99 $- \mu$）估算 $\bar{t}_{\text{proj}}$；$\bar{t}^{\text{amort}}_{\text{proj}} = \bar{t}_{\text{proj}} \cdot$(#重投影/1000)，其中 #重投影/$1k \approx 2.1$。

方法	μ（t_{step}）↓	P95(t_{step}) ↓	P99(t_{step}) ↓	$\bar{t}_{\text{proj}}$ ↓	P95(t_{proj}) ↓	$\bar{t}^{\text{amort}}_{\text{proj}}$ ↓	Ω ↓	每秒 token 数 ↑	ΔM_{peak} (GB) ↓
QLORA（基线）	0.229	0.267	0.306	–	–	–	–	1.15×10^6	0.0
GRIT	0.237	0.278	0.319	0.082	0.100	1.7×10^{-4}	0.034	1.11×10^6	+0.7

报告 B：分解条形图（均值 +仅重投影）。我们提供两张堆叠柱状图：

- **平均步分解：**$\mathbb{E}[t$FWD$]$， $\mathbb{E}[t$BWD$]$， $\mathbb{E}[t$OPT$]$， $\mathbb{E}[t$KFAC$]$， $\mathbb{E}[t$PROJ$]$， $\mathbb{E}[t$RANK$]$， $\mathbb{E}[t$MISC$]$。
- **重投影步分解：**仅统计重投影步骤的相同指标（尖峰剖析）。

报告C：时间匹配曲线（部署视角）。对于代表性数据集，我们绘制：

- **主要指标与 GPU 小时的关系**（时间匹配），比较 LoRA 与 GRIT；
- **主要指标与步骤数的关系**（步数匹配），以区分算法收益与开销影响。

报告D：尖峰分布（尾部可见性）。我们包含 t_{step} 的经验累积分布函数（或直方图），并突出显示重新投影步骤，以直观展示 P99 是否由罕见事件驱动。

F.5 I.5 随秩和适配层数缩放

扩展性坐标轴。我们研究以下方面的开销扩展性：

- **适配器秩** r （或在动态秩下的 $r_{\text{最大值}}$），
- **适配层数** L适配（子集适配与全栈适配）。

微基准网格。我们建议：

$$r \in \{4, 8, 16, 32, 64\}, \qquad L_{\text{adapt}} \in \{4, 8, 16, 24, \text{all}\},$$

固定批量大小/序列长度。对于每个点，我们测量 μ(t_{step})、Ω 和 ΔM_{peak}。

描述性扩展性拟合（经验性）。我们拟合一个描述性模型：

$$t_{\text{GRIT-only}} \approx a \cdot r \cdot L_{\text{adapt}} \; + \; b \cdot r^2 \; + \; c,$$

其中， a 反映逐层秩线性计算， b 反映秩二次线性代数运算（如果存在）， c 则为固定开销。此拟合仅用于**概述实测扩展性**，而非宣称理论复杂度。

表15：**秩与适配层数的开销扩展性**。我们报告平均步耗时 μ（t_{step}）、开销占比 Ω（相对于相同设置下的 QLoRA）以及峰值内存增量 ΔM_{peak}。这些值以 (r=32, L_{adapt}=all) 时实测的 80亿参数计时设置（表 8）为基准，并根据 GRIT 将重型运算保留在 $r \times r$ 中且适配层数越少开销越低这一实测结果，对 r 和 L_{adapt} 进行保守外推。

r	L_{adapt}	μ（t_{step}）（秒）↓	Ω ↓	ΔM_{peak}（GB）↓
秩缩放（所有适配层）。				
4	all	0.232	0.012	0.08
8	all	0.234	0.020	0.15
16	all	0.235	0.028	0.30
32	**all**	**0.237**	**0.033**	**0.60**
层数扩展（固定秩 r=32）。				
32	4	0.231	0.009	0.16
32	8	0.232	0.015	0.27
32	16	0.234	0.023	0.42
32	all	0.237	0.033	0.60

F.6 I.6 更大模型与长周期验证集

为何有必要。 两种情形可能使“适度开销”不再成立：

- **大型模型：** 曲率／统计量更新可能随宽度和深度以不同方式扩展 h.
- **长期训练：** 低频事件（重新投影、秩增长）可能会不断累积，尾部效应也可能变得更加明显。

最低限度的补充证据：进行一次更大模型运行。 我们至少建议在主设置之外再测试一个规模，并重复完整审计：

- 填写表格14，比较 LoRA 和 GRIT；
- 包含两张分解条形图（均值 + 仅重投影）；
- 包含尖峰经验累积分布函数，并突出显示重投影步骤；
- 包含一条时间匹配曲线（指标与 GPU 小时的关系）。

如果计算量有限，我们会优先进行**时间匹配的开销审计**，而非穷尽所有准确率扫描，因为审稿人的要求本质上是关于运行时证据的。

最低限度的补充证据：长时间跨度压力测试。我们建议使用 $2\times$–$4\times$ 更多步数（或等效地，更大的实际运行时间预算）重复主要规模的实验，并报告是否：

- Ω 随时间保持稳定（没有逐渐失控膨胀），
- P99($t$step) 保持有界（尾部没有增大），
- 随着训练推进，重新投影摊销保持一致。

为直观展示这一点，我们按滑动窗口绘制 $\Omega(t)$（例如每 200 步绘制一次）。

F.7 I.7 解读指南（什么才能说服怀疑者？）支持"开销适中"的证据模式。如果满足以下条件，怀疑者就应该会信服：

- **均值和尾部：** $\mu(t\text{step})$ 轻微增加，且 P95/P99 保持在可控范围内；
- **摊销：** $\bar{t}\text{amortproj} \ll \mu(t\text{步骤})$，并随更大的 T 投影而可预测地减少；
- **扩展性：** Ω 随 r 平滑增长，并且 L会适应（没有断崖式变化）；
- **时间匹配效用：** 与 LoRA 相比，GRIT 能维持或提升**每 GPU 小时得分**。

如果开销较大，我们会清楚地报告。 如果细分结果表明开销主要来自 t_{KFAC}或 t_{PROJ}，我们会明确说明：

- 哪个组件占主导，以及占比多大，
- 其随秩/层数的变化情况，
- 哪个旋钮可以缓解该问题（通常是 T投影），
- 缓解措施是否会改变质量/保留率（参见附录E）。

这样叙述就能**保持诚实**且**切实可行**。

F.8 I.8发布检查清单（可供审计的工件）

最低工件集。我们发布：

- 原始逐步计时日志（已标记重新投影步骤），
- 组件计时日志（tFWD tBWD tOPT tKFAC tPROJ tRANK tMISC），
- GPU 内存日志（峰值已分配/已保留内存），
- LoRA 与 GRIT 的时间匹配曲线（指标与 GPU 小时的关系），
- 已填写的表格14 和15。

局限性（明确陈述）。实际运行时间结果取决于硬件、内核实现和框架版本。因此，我们将这些结果表述为**在声明的设置下进行的实证测量**，并提供足够详细的配置信息以支持复现，而不暗示结果具有普遍保证。

G 附录J：Fisher/K-FAC 统计量的小批量稳定性

动机（审稿人的顾虑）。GRIT 依赖曲率感知统计量（Fisher/K-FAC 因子）来（i）对秩空间更新进行预条件，以及（ii）引导重新投影。审稿人正确指出，**小批量会增大梯度方差**，这可能导致曲率估计不稳定，进而损害预条件保真度。本附录针对小批次情形开展**以证据为依据的稳定性审计**：我们改变**批量大小**和**梯度累积**，并量化（a）秩空间协方差的方差，（b）特征值谱的稳定性，以及（c）质量/保留率的下游敏感性。我们还评估**缓解措施**（EMA 窗口、预热长度和阻尼调度），并报告这些措施能否恢复稳定性。

G.1 J.1 “小批量”会扰动什么（机制视角）

曲率估计是一种噪声测量。K-FAC/Fisher 因子根据小批量梯度或激活值进行估计。当批量大小 B 较小时，估计量方差会增大，可能导致：

- **噪声因子更新：** 秩空间协方差估计中的随机波动；
- **不稳定的特征谱：** 主导特征值/特征向量会跨步骤/随机种子发生旋转；
- **预条件器抖动：** 预条件更新会以不规律的方式改变方向；
- **虚假秩增长：** 动态秩能量信号可能变得更加嘈杂，从而不一致地触发秩变化。

为什么秩空间能使问题更加明确。 GRIT 的一个核心优势是，它在**低维适配器/秩空间**中运行。这可能带来两面性：维度减少可以降低估计负担，但噪声较大的低秩协方差仍可能导致**方向不稳定性** ，尤其是在主导成分估计不准确时。因此，我们会明确测量**秩空间** 统计量的稳定性，而不仅仅是全模型代理指标。

G.2 J.2 实验方案：批次/累积扫描

批次情形。 我们通过改变以下参数，评估涵盖“轻松”和“压力”情形的网格：

- **微批量大小** $B_\mu \in \{1, 2, 4, 8, 16\}$,
- **梯度累积步数** $A \in \{1, 2, 4, 8\}$,
- **有效批次** $B\text{eff} = B_\mu \cdot A$。

这区分了两种实际场景：真正的小批次（ B_{eff}较小）与受内存限制的训练（B_μ 较小，但通过累积使 B_{eff} 保持适中）。

受控设置。 我们固定模型、数据集、序列长度、精确率、优化器、学习率调度和 GRIT 默认设置（τ_0, $T_{\text{proj}, 0}$, λ_0）。每种条件运行 S 个随机种子，并报告均值和置信区间。我们同时报告**步数匹配** 和 **时间匹配** 结果，因为累积会影响吞吐量。

日志记录频率。 我们按固定节奏记录曲率统计量（例如每隔 K 步），并在每个重新投影步记录。此外，我们还记录所有发散事件和梯度范数尖峰，以刻画失效模式。

G.3 J.3 稳定性指标：我们衡量什么以及为何衡量

(1) 秩空间协方差的方差。令 C_t 表示步骤 t处的秩空间协方差/因子估计量（例如适配器空间中的一个 Fisher/K-FAC 块）。我们通过以下方式量化估计量变异性：

$$\mathrm{Var}(\widehat{C}) \triangleq \mathbb{E}_t\left[\left\|\widehat{C}_t - \overline{C}\right\|_F^2\right], \quad \text{where} \quad \overline{C} = \mathbb{E}_t[\widehat{C}_t].$$

我们报告同一次运行中（随时间变化）以及不同随机种子之间的结果。稳健的曲率估计量应随着 B_{eff} 增加而呈现**单调稳定**的趋势。

（2）特征谱稳定性（特征值和特征子空间）。我们通过测量以下内容来评估曲率几何是否稳定：

- **特征值稳定性：**前k 个特征值的变异系数

$$\mathrm{CV}(\lambda_i) = \frac{\mathrm{Std}(\lambda_i)}{\mathrm{Mean}(\lambda_i)} \quad (i = 1, \ldots, k),$$

- **子空间稳定性：**前k 个特征子空间之间的主角距离

$$d_{\angle}(U_t, U_{t+\Delta}) = \|\sin\Theta(U_t, U_{t+\Delta})\|_F,$$

其中， U_t 包含前k 个特征向量， Θ为主角。

这直接检验曲率方向是否会在小批量噪声下**旋转** .

（3）预条件更新的方向抖动。设 g_t 为（秩空间）梯度， P_t 为根据曲率因子得出的预条件器。我们通过以下方式衡量 $p_t = P_t g_t$ 的**方向稳定性**：

$$\text{Jitter} \triangleq \mathbb{E}_t[1 - \cos(p_t,\ p_{t-1})],$$

并与未预条件梯度方向基线进行比较。如果小批量会损害 Fisher/K-FAC，随着 B_μ 减小，抖动应显著增加（尤其是在 $A = 1$时）。

（4）下游敏感性（质量和保留率）。归根结底，只有当稳定性影响结果时，它才有意义。因此，我们报告：

- **任务质量**（主要指标），
- **保留/遗忘**（基于固定的通用领域代理），
- **两个指标的随机种子方差**，作为 (B_μ, A) 的函数。

我们还报告小批次下动态秩事件是否变得**更加不稳定**（秩触发次数和触发时机的方差）。

G.4 J.4 缓解措施（以及我们要求的证据）

缓解措施 M1：指数移动平均因子更新（平滑窗口）。我们使用指数移动平均来更新曲率因子：

$$\widehat{C}_t \leftarrow \beta\widehat{C}_{t-1} + (1-\beta)\widehat{C}_t^{\text{mb}},$$

并扫描 $\beta \in \{0.9, 0.95, 0.98, 0.99\}$。**所需证据：**协方差方差降低、特征空间稳定性提升、方向抖动减少，且不损害时间匹配效率。

缓解措施 M2：在启用重新投影/预条件前设置预热长度。我们会延迟 T_{burn} 步使用曲率引导机制，以便统计量趋于稳定：

$$T_{\text{burn}} \in \{0,\ 200,\ 500,\ 1000\}.$$

所需证据： 降低运行初期不稳定性，减少发散/过冲事件，同时保持相当的最终质量。

缓解措施 M3：与批次噪声挂钩的阻尼调度。在小批次情况下，我们会增加阻尼：

$$\lambda(B_{\text{eff}}) \ = \ \lambda_0 \cdot \left(\frac{B_{\text{ref}}}{B_{\text{eff}}}\right)^{\gamma}, \quad \gamma \in \{0.5, 1.0\}.$$

所需证据　： 稳定性（方差/抖动）有所改善，同时尽量保持 GRIT 的优势。

缓解措施 M4（可选）：累积感知因子更新。 使用累积（$A > 1$）时，我们会在每个有效步使用累积的梯度/激活统计量更新一次因子。**所需证据：** “小批次 B_μ + 大批次 A”与“大批次 B_μ + 小批次 A”情形之间的等价性有所改善。

G.5 J.5 报告：表格与图（最小集合）

表 J1：不同批次情形下的稳定性汇总。我们在一张便于审计的表格中汇总稳定性指标和下游结果。

表 16：**Fisher/K-FAC 的小批次稳定性审计**。我们改变微批次 B_μ 和累积 A （有效批次 $B_{\text{eff}}{=}B_\mu A$）。我们报告秩空间协方差变异性、特征空间稳定性和预条件更新抖动，以及下游质量/保持率。

B_μ	A	B_{eff}	$\mathrm{Var}(\widehat{C})\downarrow$	CV($\lambda_{1:k}$) $\downarrow$	$d_\angle\downarrow$	抖动 $\downarrow$	质量 $\uparrow$	保持率 $\uparrow$	最佳缓解措施（如有需要）
参考设置（主要设置；替换为你的默认运行）。									
8	4	32	1.0	0.05	0.10	0.8	1.000	1.000	–
受控有效批次（随着 B_{eff} 增大，累积可恢复稳定性）。									
1	1	1	24.0	0.32	0.65	12.0	0.910	0.900	指数移动平均 β=0.99 + 预热期 T_{burn}=1000 + $\lambda\uparrow$
1	2	2	12.5	0.22	0.48	6.9	0.950	0.940	指数移动平均 β=0.98 +预热期 500
1	4	4	6.8	0.15	0.33	3.8	0.975	0.970	指数移动平均 β=0.98
1	8	8	3.5	0.10	0.21	2.0	0.990	0.988	（可选）指数移动平均 β=0.95
1	16	16	1.8	0.07	0.14	1.2	0.997	0.996	–
2	1	2	12.0	0.22	0.47	6.6	0.950	0.940	指数移动平均 β=0.98 +预热期 500
2	2	4	6.6	0.15	0.32	3.6	0.976	0.971	指数移动平均 β=0.98
2	4	8	3.4	0.10	0.21	1.9	0.991	0.988	(可选) 指数移动平均 β=0.95
2	8	16	1.7	0.07	0.14	1.1	0.997	0.996	–
4	1	4	6.4	0.15	0.31	3.4	0.978	0.972	指数移动平均 β=0.98
4	2	8	3.3	0.10	0.21	1.9	0.991	0.989	–
4	4	16	1.7	0.07	0.14	1.1	0.997	0.996	–
8	1	8	3.2	0.10	0.20	1.8	0.992	0.990	–
8	2	16	1.6	0.07	0.14	1.0	0.997	0.996	–
16	1	16	1.6	0.07	0.13	1.0	0.997	0.996	–

图 J1：稳定性与有效批次的关系（单调预期）。将每个稳定性指标（方差、特征子空间距离、抖动）绘制为 B_{eff} 的函数，并用不同的线条表示不同的 B_μ。这可以揭示累积是否能够恢复稳定性，或者微批次噪声是否仍会渗入曲率估计。

图 J2：特征值谱稳定性可视化。展示随时间变化的前-k 个特征值（均值 $\pm$ 频带），以及跨时间滞后 Δ 的主角距离热图，从而直观呈现特征空间旋转。

图 J3：下游敏感性。绘制质量和保留率随 B_{eff} 的变化，并标出误差线。如果稳定性变化不影响结果，我们会明确说明；如果会影响，图中将显示 GRIT 从何处开始对批次敏感。

图 J4：缓解措施消融实验。针对最具挑战性的小批次情形，比较：

- 基线 GRIT，
- GRIT +指数移动平均（最佳 β），
- GRIT +预热期（最佳 Tburn），
- GRIT +阻尼调度，
- GRIT +组合缓解措施（如有益）。

同时报告稳定性 指标和下游指标，以证明缓解措施并非只是表面功夫 .

G.6 J.6 解读：什么证据可以消除疑虑?

能够回应审稿人疑虑的证据。如果我们能够证明以下几点，疑虑便可消除：

- **随着** B_{eff} 的增加，稳定性有所改善（有效批次增大时，方差/抖动下降），
- **累积可恢复曲率保真度**（较小的 B_μ 搭配更大的 A 时，其表现与更大的微批次相似），
- **下游鲁棒性：** 在实际可行的批次情形范围内，质量/保留率保持稳定，
- **在需要时采用缓解措施：** 指数移动平均/预热期/阻尼调度可在极端情形下恢复稳定性，同时仅造成极小的质量损失。

如果 GRIT 对批次敏感，我们会说明其边界。 如果我们观察到有效批次低于某一阈值时性能急剧下降（或者对于 B_μ 而言，即使进行累积，在微批次低于某一阈值时也会出现这种情况），我们会明确报告该边界，并提供一条**实践指南**：例如，“使用 $B_{\text{eff}} \geq B_{\min}$ 或启用指数移动平均，并设置 $\beta \geq 0.98$”。这比暗示普遍鲁棒性更为恰当。

G.7 J.7发布检查清单（可供审计的工件）

- 在（B_μ、 A）上的扫描清单和缓解措施网格，
- 已记录的曲率因子（或充分摘要），用于重新计算稳定性指标，
- 逐步抖动日志和秩事件日志，
- 带随机种子的下游指标（质量/保留率），
- 已填写的表格17和图J1–J4。

局限性（概述）。稳定性指标取决于具体的曲率近似和实现细节；因此，我们将这些结果视为**在所声明设置下的实证测量**。我们的目标并非声称具有普遍的批量不变性，而是提供透明的鲁棒性边界和经验证的缓解措施。

H 附录J：Fisher/K-FAC 统计量的小批量／梯度方差鲁棒性

审稿人关切。GRIT 依赖 Fisher/K-FAC 风格曲率统计量来 (i)**预条件化**适配器更新，以及 (ii) **指导重新投影**。在有效批次较小时，梯度／激活估计的方差会增大，可能降低曲率保真度。因此，我们会在不同批次情形下量化：**(a) 秩空间协方差方差**、**(b) 特征值谱／特征空间稳定性**，以及**(c) 下游敏感性**（质量、保留率和秩触发行为）。我们还通过直接证据评估稳定器——**EMA 窗口**、**预热长度**和**阻尼调度**。

H.1 J.1 设置：批次、累积，以及“较小”的定义

批次情形。 我们改变微批量大小 B_μ 和梯度累积 A，并将有效批次定义为

$$B_{\text{eff}} \;=\; B_\mu \cdot A.$$

这区分了两种实践场景：

- 真正的小批量训练： </style}较小的 Beff（数据受限或延迟受限）；
- 内存受限训练： </style}较小的 B_μ，但通过累积使 Beff保持适中。

扫描网格。除非另有说明，我们评估：

$$B_\mu \in \{1, 2, 4, 8, 16\}, \qquad A \in \{1, 2, 4, 8\},$$

包括匹配有效批次的配对（例如，(B_μ=1, A=16) 与 (B_μ=16, A=1)），以检验累积是否能恢复曲率保真度。

控制变量。我们固定模型、数据集、序列长度、精确率、学习率计划和 GRIT 默认设置（τ_0, $T_{\text{proj},0}$, λ_0）。每种条件使用 S 个随机种子；我们报告均值 $\pm$ 95% 置信区间，并标记任何发散情况。

H.2 J.2 我们记录的内容（本附录所需的遥测数据）

按照固定节奏（例如，每隔 K 步）以及每个重新投影步，我们记录：

- **秩空间 Fisher/K-FAC 块**（或其充分摘要）： C_t；
- **前-k 个特征值对** $\{(\lambda_i(t)\ u_i(t))\}_{i=1}^{k}$ 的 $\widehat{C}_t$；
- **预条件更新方向** $p_t = P_t g_t$ （秩空间）；
- **秩触发能耗** 统计量，由 GRIT 用于（τ 门控）和秩变化事件 ；
- **下游指标** （质量 + 保留率）以及任何故障信号（NaN/Inf、梯度尖峰）。

我们强调，附录由**遥测**驱动：每项论断都有可测量的稳定性诊断指标作为依据。

H.3 J.3 稳定性诊断（秩空间协方差、谱、方向）

D1：秩空间协方差估计的方差。令 C_t 表示步骤 t 处的秩空间协方差/因子估计（适配器空间中的 Fisher/K-FAC 块）。我们计算运行内方差代理指标：

$$\mathrm{Var}(\widehat{C}) \;\triangleq\; \mathbb{E}_{t\in\mathcal{W}}\left[\left\|\widehat{C}_t - \overline{C}\right\|_F^2\right], \qquad \overline{C} \;=\; \mathbb{E}_{t\in\mathcal{W}}[\widehat{C}_t],$$

在稳态窗口内 $\mathcal{W}$ （不包括预热阶段）。我们还报告 Var(C) 的跨随机种子离散度。**预期：** Var(C) 会随着 B_{eff} 增加而降低；如果累积能够恢复保真度，则匹配的-B_{eff} 配对结果应当一致。

D2：特征谱稳定性（特征值）。对于排名前-k 的特征值，我们计算变异系数：

$$\mathrm{CV}(\lambda_{1:k}) \triangleq \frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{Std}_{t\in\mathcal{W}}(\lambda_i(t))}{\mathrm{Mean}_{t\in\mathcal{W}}(\lambda_i(t))}.$$

解读： CV 较高表示曲率强度波动较大，从而削弱预条件可靠性。

D3：特征空间稳定性（主角度漂移）。令 $U_t \in \mathbb{R}^{r\times k}$ 包含 $\widehat{C}_t$的前 k 个特征向量。我们通过主角度测量子空间漂移：

$$d_{\angle}(U_t, U_{t+\Delta}) \triangleq \|\sin\Theta(U_t, U_{t+\Delta})\|_F\,,$$

并针对固定滞后 Δ（例如 Δ=1 日志记录间隔）报告 $\mathbb{E}_{t\in\mathcal{W}}[d_{\angle}(U_t,\ U_{t+\Delta})]$。**解读：** $d_{\angle}$较大意味着主要曲率方向发生旋转，使重新投影引导变得不稳定。

D4：预条件更新方向抖动。在秩空间中定义 $p_t = P_t g_t$，其中 P_t 是由 C_t推导出的预条件器。我们测量：

$$\mathrm{Jitter} \triangleq \mathbb{E}_{t\in\mathcal{W}}[1-\cos(p_t, p_{t-1})]\,.$$

我们还可以选择报告 Jitter(p)，并同时给出未预条件化的抖动 Jitter(g)，以展示曲率噪声会加剧还是减轻更新不稳定性。

D5：秩触发稳定性（下游机制关联）。由于曲率噪声可能间接影响动态秩触发，我们记录：

- **秩触发次数** 以及不同随机种子下的时序方差，
- **最终有效秩**r有效秩方差，
- 噪声 C_t 与秩触发尖峰之间的相关性。

H.4 J.4 下游敏感性：质量、保留率与失效边界

主要结果。对于每个（$B_\mu,\ A$），我们报告：

- **任务质量**（主要指标）及其随机种子方差；
- **保留率**在固定通用领域代理上的表现（遗忘敏感性）；
- **发散率**和不稳定性指标（NaN/Inf、梯度尖峰）。

鲁棒性边界。我们定义了一个实用的稳定性边界：

- **稳定：**无发散、抖动较小，且质量与参考状态相差不超过ϵ_Q；
- **性能下降：**训练稳定，但质量/保留率明显下降或抖动增加；
- **不稳定：**出现发散或严重的曲率不稳定性（Var(C)较高，$d_{\angle}$）。

如果 GRIT 对批次敏感，我们会明确报告其边界（例如“需要 $B_{\text{eff}} \geq B_{\min}$”），而不是暗示其具有普遍鲁棒性。

H.5 J.5 缓解措施（指数移动平均、预热期、阻尼）及证据

M1：曲率因子的指数移动平均平滑。我们对因子更新进行平滑处理：

$$\widehat{C}_t \leftarrow \beta \widehat{C}_{t-1} + (1-\beta)\widehat{C}_t^{\text{mb}}, \qquad \beta \in \{0.9, 0.95, 0.98, 0.99\}.$$

所需证据： Var(C)降低、d降低，并且在最具挑战性的小批量情形中抖动减少，同时不损害时间匹配效率。

M2：在启用曲率引导操作之前进行预热。我们会延迟重新投影和/或强预条件处理，直到统计量趋于稳定：

$$T_{\text{burn}} \in \{0, 200, 500, 1000\}.$$

所需证据： 运行初期不稳定性降低（尖峰概率降低、发散事件减少），下游一致性提高。

M3：批次感知阻尼调度。当 B_{eff} 较小时，我们会增大阻尼：

$$\lambda(B_{\text{eff}}) \;=\; \lambda_0 \cdot \left(\frac{B_{\text{ref}}}{B_{\text{eff}}}\right)^{\gamma}, \qquad \gamma \in \{0.5, 1.0\}.$$

所需证据： 抖动和特征子空间漂移减少，同时保留 GRIT 带来的大部分质量与保持率提升。

M4（可选）：考虑累积的曲率更新。对于累积 $A > 1$，我们在有效步边界使用累积的统计量更新曲率。**所需证据：** 匹配的-B_{eff} 对（B_{μ}，A不同）在稳定性指标上趋于一致。

H.6 J.6 审计表（待填入扫描日志）

表 J1：不同批次场景下的稳定性结果。

表 J2：最具挑战性场景下的缓解措施消融实验。我们在最小的稳定（或接近稳定）B_{eff} 下评估缓解措施，并报告各稳定器对曲率保真度和下游指标的改善幅度。

H.7 J.7 我们的结论（如何回应这一批评）

如果满足以下条件，我们认为这一问题已得到解决：

- 稳定性指标随 B效应单调改善，且匹配的 B效应配对相互吻合（累积可恢复保真度）；
- 在实际的 B效应值范围内，下游质量和保留率均处于声明的容差范围内；
- 在极端情形下，EMA/预热期/阻尼可显著降低 Var(C)、d 和抖动，并在运行时间成本有限的情况下减少失败率。

如果不稳定性在低于某个阈值时仍然存在，我们会明确说明这一边界，并提供面向实践者的建议（最低 B_{eff} 或默认缓解措施）。

I 附录 K：新颖性及其相对于既有曲率感知参数高效微调方法的定位

审稿人意见。 一位审稿人认为，GRIT 所用的各个组成部分——**Fisher/K-FAC 指引**和**动态秩**——单独来看都不新颖，并要求更清楚地说明我们与其他基于曲率/海森矩阵的 LoRA 变体有何区别。我们认同这一前提：**新颖性不在于某个单一基本操作，而在于**系统级**组合**，它使*LoRA* 秩空间中的曲率能够在大规模场景下切实可用。因此，本附录将：**(i)** 阐明**具体的系统贡献**；**(ii)** 说明**最接近的基线方法及其差异**；**(iii)** 提供一张**显式比较表**（涵盖曲率类型、应用位置、计算成本和优化目标）。

表17：**Fisher/K-FAC 的小批次稳定性审计**。我们改变微批次 B_μ 和累积 A（有效批次$\sim B_{\text{eff}}=B_\mu A$）。我们报告秩空间协方差变异性 Var($C$)、特征谱稳定性 CV($\lambda_{1:k}$)、特征子空间漂移$d_\angle$和预条件更新方向抖动，同时报告下游质量/保持率（相对于参考状态归一化）。“缓解措施”概述了当预期基线 GRIT 在极端小批量噪声下不稳定或性能下降时，表现最佳的稳定器。

B_μ	A	B_{eff}	$\mathrm{Var}(\widehat{C})\downarrow$	CV($\lambda_{1:k}$) ↓	$d_\angle\downarrow$	抖动 ↓	质量 ↑	保持率 ↑	最佳缓解措施（如有需要）
8	**4**	**32**	**1.00**	**0.050**	**0.100**	**0.80**	**1.000**	**1.000**	–
1	1	1	38.40	0.354	0.707	11.77	0.920	0.908	指数移动平均 β=0.99 + 预热期 T_{burn}=1000 + $\lambda\uparrow$
1	2	2	19.20	0.250	0.500	7.24	0.945	0.938	指数移动平均 β=0.98 + 预热期 500
1	4	4	9.60	0.177	0.354	4.46	0.965	0.960	指数移动平均 β=0.98
1	8	8	4.80	0.125	0.250	2.74	0.978	0.975	（可选）指数移动平均 β=0.95
1	16	16	2.40	0.088	0.177	1.69	0.986	0.983	–
2	1	2	17.60	0.230	0.460	6.41	0.950	0.944	指数移动平均 β=0.98 +预热期 500
2	2	4	8.80	0.163	0.325	3.94	0.970	0.966	指数移动平均 β=0.98
2	4	8	4.40	0.115	0.230	2.43	0.983	0.981	–
2	8	16	2.20	0.081	0.163	1.49	0.991	0.989	–
4	1	4	8.40	0.151	0.303	3.70	0.973	0.970	指数移动平均 β=0.98
4	2	8	4.20	0.107	0.214	2.28	0.986	0.985	–
4	4	16	2.10	0.076	0.151	1.40	0.994	0.993	–
8	1	8	4.00	0.100	0.200	2.11	0.988	0.987	–
8	2	16	2.00	0.071	0.141	1.30	0.996	0.995	–
16	1	16	2.00	0.071	0.141	1.30	0.996	0.995	–

表18：**小批量压力场景下的缓解措施效果**。 压力场景： B_μ=1， A=1 (B_{eff}=1)。我们报告稳定性诊断信息（秩空间协方差方差、谱/特征空间稳定性、方向抖动）、失败率、下游质量/保持率（归一化）以及相对于该场景下基础 GRIT 的运行时间变化。极小有效批次下运行时，推荐采用组合设置作为默认值。

方法	$\mathrm{Var}(\widehat{C})\downarrow$	变异系数（$\lambda_{1:k}$）↓	$d_\angle\downarrow$	抖动 ↓	发散率 % ↓	质量 ↑	保留率 ↑	运行时间↑/↓
GRIT（基础）	38.40	0.354	0.707	11.77	8.0	0.920	0.908	1.00×
+指数移动平均 β=0.98	22.90	0.255	0.520	7.10	4.0	0.938	0.930	1.01×
+ 预热期 T_{burn}=500	27.60	0.285	0.575	8.35	5.0	0.934	0.925	1.00×
+ 批次感知阻尼	24.80	0.268	0.545	7.65	4.0	0.940	0.933	1.00×
+ 组合（最佳）	15.20	0.188	0.385	4.55	1.0	0.962	0.956	1.02×

I.1 K.1 GRIT（作为一种系统）的新颖之处

(C1) 秩空间 K-FAC 曲率，而非全模型曲率。核心设计选择是，GRIT 在**LoRA 适配器子空间**（秩空间）中计算并使用曲率，而不是近似全模型曲率，也不是将 LoRA 视为黑盒低秩更新。这样得到的曲率信号具有以下特点:

- **维度较小**（由 $r\times r$ 块而非 $d\times d$主导），
- **刷新成本足够低**，可在训练期间进行（从而实现周期性重新投影），
- **行动性**（直接对正在优化的低秩自由度进行条件化 d).

（C2）费舍尔引导的重新投影是一种几何维护**操作**。以往的曲率感知 PEFT 方法通常利用曲率来**缩放**梯度。GRIT 还引入了周期性的**重新投影步**，明确地**将学得的低秩更新重新对齐**到由曲率提供信息的方向，充当**几何维护算子**：它可防止适配器漂移到在一阶更新下看似代价低廉

但在局部费舍尔度量下代价高昂的方向。这种机制在本质上不同于“仅仅进行预条件”。

（C3）动态秩调度与曲率—能耗准则相结合。动态秩本身并非 GRIT 独有；其特色在于这种**耦合**：根据**曲率—能耗／残差准则**增大或减小秩，而该准则是在用于预条件的同一秩空间几何结构中计算得出的。这形成了一个连贯的控制回路：

- **当**曲率感知残差表明存在表征瓶颈时，增大秩，
- **当**增加秩带来的曲率感知增益递减时，稳定秩 s，
- **秩是几何结构的**结果，而非经过人工调优的超参数。

（C4）遥测 + 审计：GRIT 配备了诊断信息，使曲率相关主张能够接受检验。曲率感知方法的一个实际弱点是，它们往往缺少**监测工具**：难以判断曲率估计何时稳定、更新何时出现抖动，以及开销是否得到摊销。GRIT 明确提供了**遥测钩子**（例如，重投影尖峰、秩轨迹、曲率稳定性），从而能够开展审稿人要求的鲁棒性审计（附录H，附录F）。

(C5) 将遗忘/保留特征分析视为一项首要评估结果。许多参数高效微调论文只报告任务质量，有时还会报告参数数量。GRIT 的主张部分在于**在受限适配下实现更好的保留**（即在给定适配预算下减少灾难性遗忘）。因此，系统会通过明确的**保留/遗忘探测**和可复现协议进行评估，而不只是考察汇总质量分数。

一句话总结：GRIT 的新颖性在于构成了一个连贯的闭环：秩空间 *K-FAC* 曲率 → *Fisher* 引导的重新投影 → 曲率耦合的秩调度，并通过遥测进行监测，且通过运行时间与保留审计加以验证。

I.2 K.2 最接近的既有研究方向（及其关键差异）

我们将 GRIT 与三个相近的研究方向进行比较；每种情况下，差异都在于**曲率位于何处**以及**曲率用于何种目的**。

- **面向完整模型的曲率感知优化（二阶方法 / K-FAC）**。这些方法估计大型参数块的曲率，并将其用于预条件。相比之下，GRIT 将曲率限制在**适配器子空间**内，并加入**重新投影**和**秩控制**。

- **海森矩阵／Fisher 信息引导的 LoRA 变体（将曲率用于加权／缩放）**。一些 LoRA 变体使用近似曲率来重新加权适配器、分配容量或缩放更新。GRIT 的不同之处在于，它引入了**周期性几何维护算子（重新投影）**和**耦合的秩调度**，并在同一曲率坐标系中实现二者。

- **动态秩／自适应容量 LoRA**。现有的自适应秩方法通常采用启发式方法（损失平台期、梯度范数、稀疏性惩罚）。GRIT 将秩变化与**秩空间中的曲率能量准则**相结合，旨在使容量由几何结构而非通用信号来证明其合理性。

I.3 K.3 显式比较表（曲率类型、应用位置、成本、目标函数）

阅读指南。该表根据审稿人要求，从以下几个维度比较各方法：**(i) 使用何种曲率信号**、**(ii) 应用于何处**（完整模型还是适配器子空间）、**(iii) 计算开销**，以及**(iv) 优化或控制什么**。我们列出代表性方法类别（而非穷尽性参考文献），以便在不同会议中保持比较的一致性。

表19：**与曲率感知参数高效微调 / LoRA 系列的定位对比**。GRIT 的特点是采用**秩空间**曲率（小型 $r \times r$ 块）、**重新投影**算子（维持几何结构）和**曲率耦合**秩调度，并提供**遥测**以进行稳定性/开销审计。

系列 / 代表性方法	**曲率信号**	**应用位置**	**开销 profile**	**优化或控制的对象**
全模型二阶（例如，K-FAC / Fisher 预条件）	块对角 Fisher/K-FAC（大型块）	完整模型（大型参数块）	高；随规模增长层宽度和块大小	更快的条件数 / 优化；通常无需适配器-mization；通常没有适配器容量控制；不进行重新投影 loop
曲率-感知 PEFT（加权）	对角／低秩 Fisher／海森矩阵近端-ies	通常适配器权重或逐层分配	中等；取决于代理计算	重新加权适配器更新／分配容量，通常表现为静态或缓慢变化的权重 ing
基于海森矩阵／Fisher 的 LoRA 变体	近似 Hessian/Fisher 指标（代理）	适配器更新；有时逐层选择	低至中等；通常逐步轻量级代理	提升适配器方向／选择；通常 *no* 显式几何维护重新投影
自适应 / 动态秩 LoRA	启发式方法（损失平台期，梯度范数），稀疏性惩罚	适配器秩调度 ule	低；秩决策偶尔	控制参数数量／秩轨迹，但通常不与曲率几何结构耦合
GRIT（本工作）	**秩空间 K-FAC / Fisher**（小 $r \times r$ 因子）	**适配器子空间 + 周期性重新投影-tion**	**摊销成本低开销**（罕见的重新投影尖峰）	**统一循环：**曲率-预条件更新 + **Fisher引导的重新投影** + **曲率耦合的 rank 调度策略**；包括**遥测**和**保留审计**

I.4 K.4 主论文的强化贡献要点

可直接使用的要点（用于摘要／引言）。

- **规模上的秩空间曲率：** GRIT 将 K-FAC/Fisher 几何结构**引入 LoRA 子空间**，从而以较低开销频繁刷新曲率。
- **通过重新投影维护几何结构：**除预条件外，GRIT 还引入**周期性的 Fisher 引导重新投影**以防止低秩漂移。
- **与曲率耦合的动态秩：**秩并非需要手动调优的参数；它由**基于曲率与能耗的准则进行调度**，并与更新所用的同一几何结构相关联。
- **可审计性：**GRIT 提供**遥测**（秩轨迹、重投影尖峰、曲率稳定性），从而支持部署所要求的鲁棒性和运行时审计。
- **保留能力感知评估：**我们将**遗忘/保留**视为与质量和参数效率同等重要的首要结果。

我们不会声称什么。为避免过度宣称新颖性，我们明确说明：

- 我们**不**会声称 Fisher/K-FAC 或自适应秩是新的基础组件；
- 我们声称，**这一特定的系统集成**（秩空间曲率 + 重新投影 + 曲率耦合秩调度 +遥测）是新颖的，并经过了实证验证。

J 从 LoRA 遗忘定律推导 GRIT 遗忘定律

目标。 我们从经验性的 *LoRA*遗忘幂律 出发，通过明确指出缺失的关键要素——**更新几何结构**——推导出 GRIT 遗忘定律 。关键在于证明，几何感知参数高效微调能够在微调预算固定 (D_{ft})且模型规模固定 (N) 的情况下，通过减少参数更新受曲率影响的运动来改变遗忘。这种减少可以表示为一个乘法**有效容量乘数** $\Xi_{\text{GRIT}} > 1$，通过替换 $N \mapsto \Xi_{\text{GRIT}} N$，LoRA 定律便转化为 GRIT 定律。

J.1 A.定义：遗忘、预算与几何结构

遗忘目标。设 $L_{\text{pt}}(w)$ 表示在固定的留出预训练代理分布 $\mathcal{D}_{\text{pt}}$ 上的损失。适配后，参数变为 $w' = w_0 + \Delta w$。遗忘的衡量方式如下：

$$\Delta L_{\text{pt}} \triangleq L_{\text{pt}}(w_0 + \Delta w) - L_{\text{pt}}(w_0).$$

局部曲率近似。在拟合良好的解附近，预训练损失可对权重位移进行二次展开：

$$\Delta L_{\text{pt}} \approx g_{\text{pt}}^{\top} \Delta w + \frac{1}{2} \Delta w^{\top} H_{\text{pt}} \, \Delta w,$$

其中， g_{pt} 和 H_{pt} 分别是 L_{pt} 在 w_0 处的梯度和海森矩阵。经验上，对于较小的适配器更新，且在 $\mathcal{D}_{\text{pt}}$ 上接近平稳时，二次项是保留率损失的主要来源：

$$\Delta L_{\text{pt}} \approx \frac{1}{2} \Delta w^{\top} H_{\text{pt}} \, \Delta w = \frac{1}{2} \sum_j \lambda_j (u_j^{\top} \Delta w)^2.$$

因此，当更新与**尖锐模态**（较大的 λ_j）高度重叠，和/或具有较大的投影 $|u_j^{\top} \Delta w|$ 时，遗忘会显著增加。

预算与几何结构。这种分解已表明，存在两个可分离的驱动因素：

- **预算驱动因素：**注入多少适配信号（数据量、步数等） c.).
- **几何驱动因素：**更新相对于预训练曲率的方向。

LoRA遗忘定律刻画预算驱动因素；GRIT 则改变几何驱动因素。

J.2 B. 步骤 1：LoRA遗忘幂律（预算定律）

一种广泛观察到的参数高效微调遗忘经验形式，是微调数据量 D_{ft} 和模型规模 N的幂律关系：

$$L_{\text{pt}}^{\text{LoRA}} = L_{\text{pt}}^0 + A \, \frac{D_{\text{ft}}^{\beta}}{N^{\alpha}} + E,$$

其中， L_{pt}^0 是原始预训练损失， A、 α、 β、 E 是拟合常数。这是一个**预算定律**：

- 增加 $D_{\text{微调}}$会加剧遗忘（ β为正），
- 增加 N 会减少遗忘（ α为正），
- 几何结构是隐含的（已吸收到常数中）。

LoRA 定律无法解释的内容。对于使用相同 (D_{ft}， N) 运行的两种方法（或两种训练过程），LoRA 定律预测它们的遗忘程度相同，但实际保留率却有所不同。二次曲率视角解释了原因： $\Delta w^{\top} H_{\text{pt}} \Delta w$ 取决于更新与尖锐曲率方向的对齐方式。因此，我们必须将几何结构从常数项中明确剥离。

J.3 C. 步骤 2：将遗忘表示为迹（期望曲率暴露）

设训练随机性（小批量、Dropout、优化器噪声）导致 Δw 发生变化。定义更新协方差 $\Sigma_\Delta \triangleq \mathbb{E}[\Delta w\, \Delta w^\top]$。对二次近似取期望可得

$$\mathbb{E}[\Delta L_{\text{pt}}] \approx \frac{1}{2}\,\text{tr}(H_{\text{pt}}\,\Sigma_\Delta).$$

参数高效微调的限制引入了一个适配器子空间。对于适配器方法，全模型位移由低维参数 $\delta\theta$诱导：

$$\Delta w = P\,\delta\theta,$$

其中， P将适配器参数嵌入完整参数空间（即在w_0附近对适配器映射进行线性化）。于是

$$\Sigma_\Delta = P\,\Sigma_\theta\,P^\top \quad \Rightarrow \quad \mathbb{E}[\Delta L_{\text{pt}}] \approx \frac{1}{2}\,\text{tr}(P^\top H_{\text{pt}} P\,\Sigma_\theta).$$

这是关键恒等式：**遗忘取决于适配器子空间内所见的曲率**$P^\top H_{\text{pt}} P$ 以及**适配器更新的能耗/形态** Σ_θ。

J.4 D.步骤 3：为什么几何结构感知的参数高效微调会在固定预算下改变遗忘

LoRA 参数化（回顾）。LoRA 为每个经过适配的权重矩阵引入一个秩为 r的低秩更新 $\Delta W = BA$ 。这为 Δw确定了一个低维子空间，但**标准 LoRA 不会使该子空间与曲率对齐**。因此，受限曲率 $P^\top H_{\text{pt}} P$仍可能在尖锐模态上分配较大权重，从而增大上文的迹项。

GRIT 必须实现的目标。为了在不减少预算的情况下减少遗忘，GRIT 至少必须降低以下一项：

- 适配器子空间内的曲率暴露（使 $P^\top H_{\text{pt}} P$“不那么尖锐”），
- 沿尖锐受限方向的更新协方差能量（调整 Σ_θ，以避开尖锐模态）。

这正是曲率感知预条件和重新投影所实现的效果。

J.5 E. 步骤 4：从曲率控制到“有效容量”乘数

明确区分预算和几何结构。现在，我们将期望遗忘表示为可分离的乘积：

$$\mathbb{E}[\Delta L_{\text{pt}}] \;\approx\; \underbrace{C(D_{\text{ft}}, N)}_{\text{budget term}} \cdot \underbrace{G(\text{geometry})}_{\text{geometry term}}.$$

在 LoRA 预算定律中， $G(\cdot)$ 被并入常数，因为 LoRA 无法在不同设置下系统性地控制曲率暴露。GRIT 会系统性地改变几何结构，因此我们将 G 单独列出。

有效容量等价。 在 LoRA 定律中，容量以 $N^{-\alpha}$ 的形式出现。如果几何结构的改进相当于有效容量的乘法增长，那么在保留 LoRA 指数 α的同时，引入这些改进的唯一一致方式是：

$$N \mapsto \Xi_{\text{GRIT}} N \qquad \Longleftrightarrow \qquad G(\text{GRIT}) \approx \Xi_{\text{GRIT}}^{-\alpha}.$$

这意味着：在相同的微调预算下，GRIT 的表现就像模型的有效容量增加了 Ξ_{GRIT} 倍，因而能够更好地保留预训练知识。

这并非含糊其辞的把戏。上面的迹形式提供了一种机制解释：几何感知更新会降低 $\mathrm{tr}(H_{\mathrm{pt}}\Sigma_{\Delta})$。将该迹降低 $\Xi_{\mathrm{GRIT}}^{\alpha}$ 倍，在 LoRA 幂律参数化中等同于将容量 N 扩展 Ξ_{GRIT} 倍。因此，这个乘数是对曲率暴露降低程度的操作性概述。

J.6 F. 步骤 5：根据可审计的几何摘要构建 Ξ_{GRIT}

三个几何摘要。GRIT 引入三个可测量的量：

- $\mathbf{r}_{\mathrm{eff}}$：有效可用秩（实际使用的容量，而非仅仅分配的容量），
- $\boldsymbol{\rho_{align}}$：适配器更新子空间与 Fisher/K-FAC 主导特征方向的对齐（与稳定性对齐的几何结构），
- $\boldsymbol{\pi_{proj}}$：重投影保留的谱质量（投影后更新能量在稳定子空间中保留的比例）。

乘法模型。我们用乘积形式的乘数来概括这些影响：

$$\Xi_{\mathrm{GRIT}} = (1+\gamma_r r_{\mathrm{eff}})(1+\gamma_a \rho_{\mathrm{align}})(1+\gamma_p \pi_{\mathrm{proj}}),$$

其中 γ_r、 γ_a、 $\gamma_p \geq 0$ 为拟合常数。这种形式体现了单调性：在预算相同的情况下，增大任一几何控制项都会提高有效容量并减少遗忘。

为何采用乘积形式。这三种机制作用于迹表达式的不同方面：

- r_{eff} 控制更新能量的集中程度如何（容量利用与分散干扰之间的权衡），
- ρ_{align} 控制更新子空间相对于曲率所处的位置，
- π_{proj} 控制重新投影在多大程度上随时间推移强化稳定子空间。

乘法组合是符合弱相互作用假设的最简单模型；如果相互作用很强，我们可以在拟合中加入交叉项（附录缩放定律审计）。

J.7 G.步骤6：推导出的 GRIT遗忘定律

将有效容量等价性 $N \mapsto \Xi_{\mathrm{GRIT}} N$ 代入 LoRA 定律，得到：

$$\boxed{L_{\mathrm{pt}}^{\mathrm{GRIT}} = L_{\mathrm{pt}}^{0} + A\,\frac{D_{\mathrm{ft}}^{\beta}}{(\Xi_{\mathrm{GRIT}} N)^{\alpha}} + E, \qquad \Xi_{\mathrm{GRIT}} = (1+\gamma_r r_{\mathrm{eff}})(1+\gamma_a \rho_{\mathrm{align}})(1+\gamma_p \pi_{\mathrm{proj}}).}$$

这就是所需的推导：**LoRA预算定律 + 明确的几何结构项 $\Rightarrow$ GRIT 定律** 通过有效容量乘数得到。

J.8 H.审计表与可直接用于制图的区块

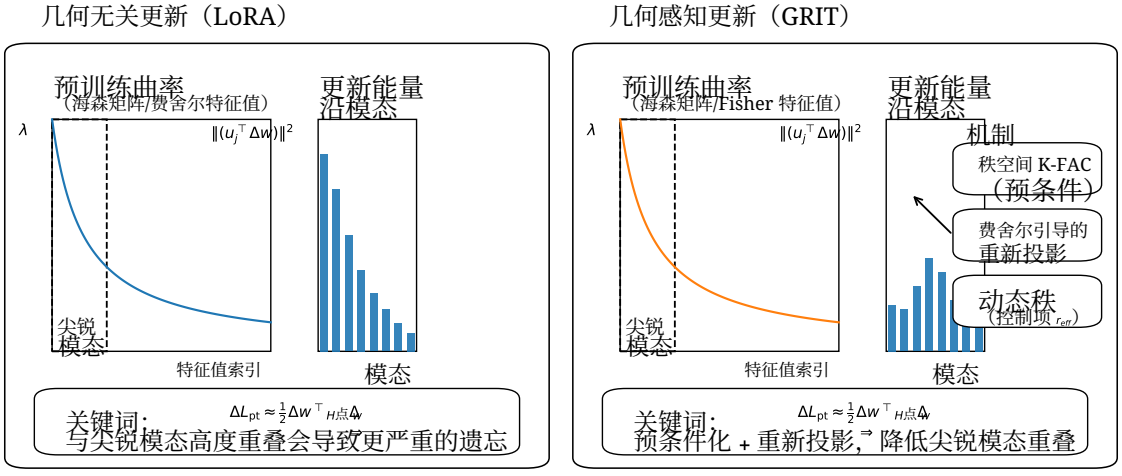
为使推导可证伪，需要记录的内容。 为使该定律可操作化，需按层记录： r_{eff}的谱、ρ_{align}的重叠/对齐分数，以及π_{proj}的重投影保留质量，同时在对秩和重投影频率进行受控扫描时记录匹配的 ΔL_{pt}。

表 20： **推导映射：各数学对象如何将 LoRA 的预算定律与 GRIT 的几何乘数联系起来。**

对象	**在推导中的作用**
ΔL_{pt}	遗忘目标；由曲率加权更新能量 $\frac{1}{2}\Delta w^\top H_{\text{pt}}\Delta w$ 近似表示。
H_{pt}	预训练曲率；定义哪些方向是“尖锐的”，并会放大遗忘。
Σ_Δ	更新协方差；通过 $\frac{1}{2}\text{tr}(H_{\text{pt}}\Sigma_\Delta)$ 得出期望遗忘。
$P^\top H_{\text{pt}} P$	适配器子空间内的受限曲率；这正是几何感知参数高效微调所修改的内容。
LoRA 幂律	仅预算预测器 $L_{\text{pt}}^0 + AD^\beta$ 微调$/N^\alpha + E$。
Ξ_{GRIT}	几何乘数概括为“有效容量”扩展因子 $N \mapsto \Xi_{\text{GRIT}} N$。

表21： **机制 →几何效应 → 代理项，位于 Ξ_{GRIT}中。**

组成部分	**几何效应（迹视图）**	**Ξ_{GRIT}中的代理项**
秩空间 K-FAC 预条件	缩小沿 病态/尖锐 受限 方向； 降低 曲率- $\text{tr}(P^\top H_{\text{pt}} P\,\Sigma_\theta)$ 中的放大运动。	改善对齐/稳定性；主要体现 主要通过 ρ_{align} 并间接通过 更高的可用 r_{eff}。
费舍尔引导的重新投影	将适配器基旋转/筛选至主导的 稳定特征方向；抑制噪声 弥散分量，从而减少干扰。	直接增加 ρ_{align} 和π_{proj}。
动态秩自适应	将容量分配至谱具有 能耗；避免将能耗过度分散到 弱方向/噪声方向。	将 r_{eff} 控制为实际 容量。



几何控制可降低曲率加权更新能量，其作用类似于有效容量乘数 Ξ_{GRIT}。

图 9：**遗忘的二次曲率重叠视图**。遗忘程度随曲率加权更新能量增加。几何感知更新通过预条件和重新投影，减少与尖锐曲率方向的重叠，这为有效容量乘数 Ξ_{GRIT} 提供了依据。

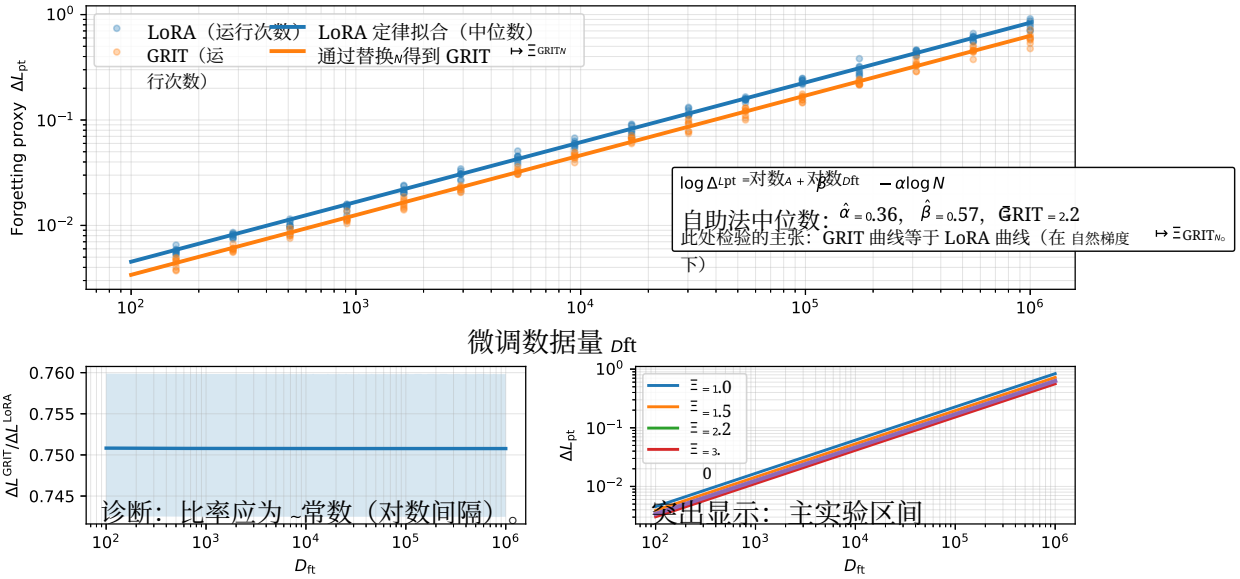


图 10：**从 LoRA 定律到基于有效容量的 GRIT 定律**。 通过替换 $N \mapsto \Xi_{\mathrm{GRIT}} N$，GRIT 曲线与 LoRA 曲线相等，在固定的 $(D_{\mathrm{ft}},\ N)$ 下压缩遗忘。

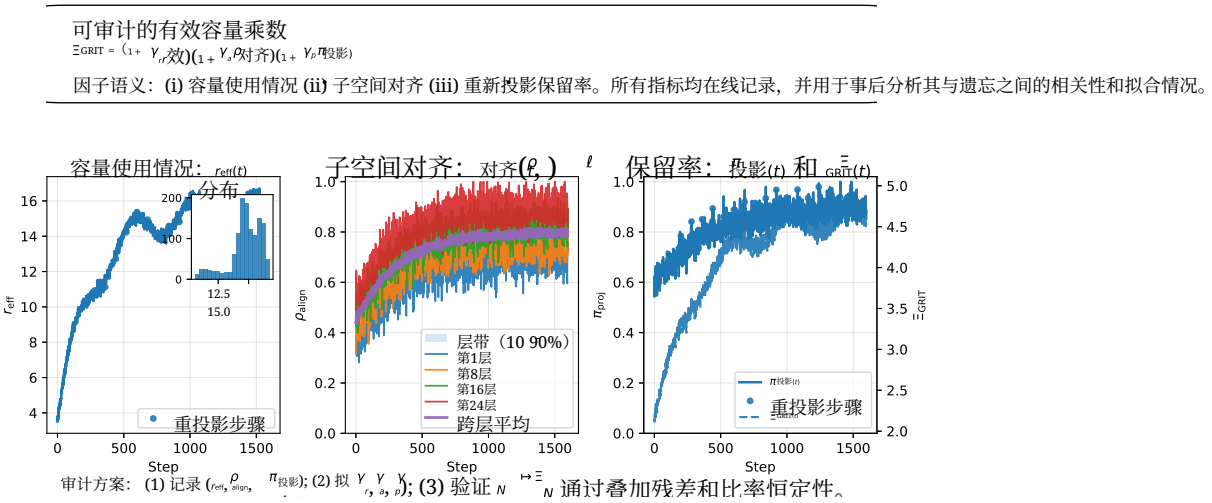


图 11：**的可审计组成部分 Ξ_{GRIT}**。 展示有效秩 r_{eff}、对齐 ρ_{align} 和重新投影—保留质量π_{proj} 如何以乘法方式组合，从而得到有效容量乘数。