Norm Calibration and Soft Centering Reveal Two Types of Seed-Dependent Collapse in LoRA Fine-Tuning

Anonymous ACL submission

Abstract

LoRA fine-tuning can suffer from seed-dependent collapse: on certain random seeds, the adapted model falls to near-chance accuracy on specific benchmarks while training and validation loss remain unremarkable. Studying this on eight commonsense benchmarks with LLaMA 3-8B, we identify two qualitatively distinct collapse regimes and propose the BNC family to address them. BNC (**B**ounded **N**orm-**C**alibrated LoRA) resolves *amplitude-driven collapse* with negligible overhead by clipping each token's update to a ν-fraction of the base output norm (HellaSwag seed 123: 54.7% $\rightarrow$ 94.2%) and improves monotonically across the tested ranks. BNC-C (**C**entered BNC) extends BNC with a soft per-token centering step parameterised by β, resolving *boundary-approximate collapse*: at rank 8 ($\beta = 0.75$), BoolQ seed 123 recovers from 62.3% to 74.9%. With calibrated β, BNC-C reaches the best overall mean we observe ($85.5\% \pm 0.3\%$ at rank 32), while BNC remains the zero-configuration default when per-rank tuning is not desired. Centering strength is rank-sensitive: the failed $r = 16$, $\beta = 0.75$ checkpoints cannot be repaired by eval-time overrides, indicating that centering acts through the training trajectory rather than as a post-hoc inference correction.

1 Introduction

Parameter-efficient fine-tuning (PEFT) keeps pre-trained weights frozen and injects a small number of trainable parameters. LoRA (Hu et al., 2022) decomposes the weight update as $\Delta W = BA$, scaled by the fixed factor α/r. Despite its widespread adoption, LoRA exhibits seed-dependent collapse: on certain random seeds the adapted model collapses to near-chance accuracy on specific downstream benchmarks, even while training loss and validation loss follow apparently normal trajectories.

A key open question is whether all observed collapse instances share the same underlying cause. If not, different collapse types may require different solutions. Existing stabilisation strategies—such as learnable scaling (Kalajdzievski, 2023), label smoothing (Müller et al., 2019), and the coupled-gate architectures explored in our DyS variants—each address one axis of the problem but do not characterise whether they target the same failure mode.

We propose BNC (**B**ounded **N**orm-**C**alibrated LoRA), a per-token norm constraint that limits LoRA updates from overwhelming base-model representations:

$$h_i = W_{0,i}\, x + \gamma_m \cdot \mathrm{clip}_\nu(B_i A_i\, x;\ W_{0,i}\, x)\,, \quad (1)$$

where $\mathrm{clip}_\nu(u;\, s) = u \cdot \min(1,\, \nu\, \|s\|_2 / (\|u\|_2 + \varepsilon))$, γ_m is a learnable per-module gate initialised to α/r and bounded within $\pm 20\%$ of α/r, applied independently to each token. BNC adds one learnable scalar gate per LoRA module (γ_m), amounting to 160 scalars for LLaMA 3-8B ($< 0.001\%$ of total LoRA parameters); the norm ratio ν is the only hyperparameter.

Applying BNC and inspecting same-rank ablations reveals that the seed-dependent collapses we observe fall into two qualitatively distinct categories. One is repaired by directly constraining update magnitude; the other requires additional control over update geometry at the same rank.

Amplitude-driven collapse. In this type, the LoRA update magnitude spirals to values that completely override the base-model output on a specific task-seed combination. HellaSwag at seed 123 (rank 8) is the clearest case: LoRA collapses to 54.7% (near chance for a 4-way task), while BNC restores it to 94.2%. This collapse responds to interventions that directly limit update magnitude or prediction overconfidence; among the methods we evaluate, BNC provides the cleanest recovery.

Boundary-approximate collapse. In this type, the LoRA update accumulates a mean-direction bias that dominates the binary decision boundary at certain seeds. BoolQ at seed 123 (rank 8) is the clearest case: LoRA, DoRA, and BNC all produce near-identical collapsed BoolQ accuracy ($\approx 62\%$), while standard training loss and validation loss show no sign of failure. This regime is not a hard capacity wall: BNC at rank 16 reaches 75.0%, and both TAC and tighter BNC ($\nu = 0.5$) recover BoolQ seed 123 at rank 8. Mechanistic ablations within the tested UGC presets suggest that mean-direction centering is the main distinguishing ingredient: trust-only and single-component variants plateau at 66.8%, while only full TAC crosses that plateau within the UGC family.

We introduce BNC-C (**C**entered BNC), which applies soft centering directly within the BNC pipeline and forms the second member of the BNC family. Setting $\beta = 0$ recovers BNC exactly; BNC-C is thus a strict extension of BNC. At $\beta = 0.75$, BNC-C resolves both observed collapse types in the rank-8 setting: BoolQ seed 123 recovers from 62.3% to 74.9%, HellaSwag remains stable around 95.4% across all three canonical seeds, and overall cross-seed variance drops from $\pm 1.0\%$ to $\pm 0.1\%$.

Contributions.

1. We introduce a collapse taxonomy distinguishing *amplitude-driven* and *boundary-approximate* collapse, characterised by their differential responses to norm calibration, rank variation, and update geometry. A key diagnostic finding is that the representative collapsed runs we inspect are not reliably signalled by standard optimisation metrics; the collapsed seed can even achieve validation loss comparable to or lower than stable seeds.

2. We propose BNC, a per-token bounded norm-calibrated LoRA that prevents the observed amplitude-driven collapse by bounding each token's adapter update to a fraction of the base-model output norm. It adds one learnable scalar gate per LoRA module (160 scalars for LLaMA 3-8B, $< 0.001\%$ of total LoRA parameters).

3. We provide multi-rank evidence ($r \in \{8, 16, 32\}$) that BNC is a robust cross-rank stabiliser, achieving $84.4\% \pm 1.0\%$, $84.8\% \pm 0.7\%$, and $85.2\% \pm 0.1\%$ with variance shrinking monotonically. We further propose BNC-C, which augments BNC with soft per-token centering and a learnable gate. With $\beta = 0.75$, it improves over BNC at $r = 8$ ($84.9\% \pm 0.1\%$, BoolQ std $6.8\% \rightarrow 0.7\%$) and $r = 32$ ($85.5\% \pm 0.3\%$). We show that centering strength is rank-sensitive: $\beta = 0.75$ over-corrects at $r = 16$ ($83.6\% \pm 1.4\%$, training-induced), but reducing to $\beta = 0.5$ recovers this rank to $85.1\% \pm 0.04\%$, above BNC. Eval-time β overrides cannot repair the failed checkpoints, confirming the effect is a training-time update-shaping mechanism.

4. Via controlled UGC ablations, we show that mean-direction centering is the main distinguishing component among the tested presets for recovering boundary-approximate collapse: simpler geometry variants plateau at 66.8% on BoolQ seed 123, while TAC and BNC-C exceed that plateau.

2 Background

Low-Rank Adaptation (LoRA). Given a pre-trained weight $W_0 \in \mathbb{R}^{d \times k}$, LoRA parameterises:

$$h = W_0 x + \frac{\alpha}{r} BAx, \tag{2}$$

where $A \in \mathbb{R}^{r \times k}$, $B \in \mathbb{R}^{d \times r}$, $r \ll \min(d, k)$, and α/r is a fixed hyperparameter.

Weight-Decomposed Low-Rank Adaptation (DoRA). DoRA (Liu et al., 2024) decomposes W_0 into magnitude and direction components and applies low-rank updates to the direction. It can improve accuracy over LoRA but is slower due to per-step column-norm recomputation.

Rank-Stabilised LoRA (rsLoRA). rsLoRA (Kalajdzievski, 2023) replaces the fixed scalar with $\alpha/\sqrt{r}$, improving training stability across rank settings. The scalar remains fixed throughout training.

3 Method

3.1 BNC: Bounded Norm-Calibrated LoRA

Standard LoRA applies a fixed global scaling factor α/r to the low-rank update. When this update's magnitude becomes large relative to the base-model output, the adapter can override the pre-trained representation and produce highly seed-dependent predictions.

BNC addresses this by constraining each token's LoRA update to at most a ν-fraction of the corresponding base-model output norm. For a single token representation with base output $s = W_0 x$ and LoRA update $u = \frac{\alpha}{r} BAx$ (absorbing the standard LoRA scaling into u):

$$\mathrm{clip}_\nu(u;\, s) = u \cdot \min\left(1,\ \frac{\nu\, \|s\|_2}{\|u\|_2 + \varepsilon}\right), \qquad (3)$$

with $\varepsilon = 10^{-6}$. The forward pass is:

$$h = W_0 x + \gamma_m \cdot \mathrm{clip}_\nu(BAx;\, W_0 x)\,. \qquad (4)$$

Properties. Unlike fixed-scalar methods (LoRA, rsLoRA), the effective clipping threshold adapts to each token's base output magnitude—tokens with stronger base activations receive proportionally larger maximum updates. Unlike gate-heavy methods (DyS-GS-Res, DyS-GS), BNC adds only 160 scalar gate logits ($< 0.001\%$ of total parameters). The clip operator is piecewise differentiable almost everywhere and leaves the standard LoRA gradients unchanged when it is inactive. Bounding γ_m near the standard LoRA scaling keeps the learned gate from undoing the norm cap while still allowing modest per-module adaptation. The hyperparameter ν is set before training; unless otherwise noted, BNC uses $\nu = 1.0$, and we additionally report a $\nu = 0.5$ ablation for the fragile BoolQ rank-8 setting.

3.2 BNC-C: Centered Bounded Norm-Calibrated LoRA

BNC eliminates the observed amplitude-driven collapse but leaves BoolQ at seed 123 near the majority baseline (62.3%), identical to standard LoRA. Inspection of the predicted-"true" ratio reveals that the LoRA update on this seed develops a strong mean-direction bias: the per-token row mean $\bar{u} = \frac{1}{d}\sum_j u_j$ dominates the update, pushing all predictions toward one answer. This bias is seed-specific—seed 123 on BoolQ accumulates it while other seeds and tasks do not.

BNC-C corrects this by subtracting a fraction β of the hidden-dimension mean before applying the norm clip:

$$\tilde{u} = u - \beta\, \bar{u}\, \mathbf{1}, \qquad (5)$$

$$h = W_0 x + \gamma_m \cdot \mathrm{clip}_\nu(\tilde{u};\, W_0 x), \qquad (6)$$

where $u = BAx$ is the raw token update, $\bar{u} = \mathrm{mean}(u, \mathrm{dim}{=}{-1}, \mathrm{keepdim}{=}\mathrm{True})$ is its hidden-dimension mean, $\mathbf{1}$ broadcasts that scalar mean back across the hidden dimension, γ_m is a learnable per-module gate initialised to α/r and bounded within $\pm 20\%$ of α/r, and clip_ν is the same per-token norm clip as in BNC (Eq. 3). Setting $\beta = 0$ removes the centering step ($\tilde{u} = BAx$), making Eq. 6 identical to Eq. 4: BNC-C with $\beta = 0$ recovers BNC exactly. This preserves BNC's norm-calibrated update path while adding only one extra centering coefficient, making BNC-C a lighter-weight geometry correction than the full TAC pipeline.

Centering strength ablation. We test $\beta \in \{0.25, 0.50, 0.75\}$ at rank 8 on seeds 42 and 123. $\beta = 0.25$ is too weak: BoolQ seed 123 remains at 62.2%, unchanged from BNC. $\beta = 0.50$ causes broad training instability on seed 123 (HellaSwag drops to 83.8%, ARC-Challenge to 63.0%), indicating that this centering strength destabilizes optimization without recovering BoolQ. $\beta = 0.75$ is the first setting that crosses the recovery threshold: BoolQ seed 123 rises to 74.9% while HellaSwag remains at 95.4%—matching the level achieved by TAC while preserving stable performance on the three canonical seeds (§4.3). Figure 2 visualises the full sweep. Unless otherwise noted, BNC-C uses $\beta = 0.75$ at $r \in \{8, 32\}$ and $\beta = 0.5$ at $r = 16$, where $\beta = 0.75$ over-corrects (see §4.5).

Relationship to TAC. We use TAC as a reference geometry-controlled update pipeline rather than as a co-equal primary method. TAC applies full centering followed by shape control and trust-region clipping, whereas BNC-C uses the same centering idea inside the simpler BNC pipeline. Empirically, this variant preserves stable performance on the three canonical seeds while retaining the main recovery benefit of TAC.

Mechanism probes. We use TAC and simpler geometry-controlled presets only as mechanism probes for the rank-8 BoolQ plateau. Among the tested same-rank variants, only full TAC crosses the observed partial-recovery threshold; trust-only and simpler centered variants remain below it (Table 2). We therefore treat TAC as supporting evidence that update geometry matters, rather than as a co-equal primary method.

4 Experiments

4.1 Setup

Base model. LLaMA 3-8B (Meta AI, 2024) with 32 transformer layers.

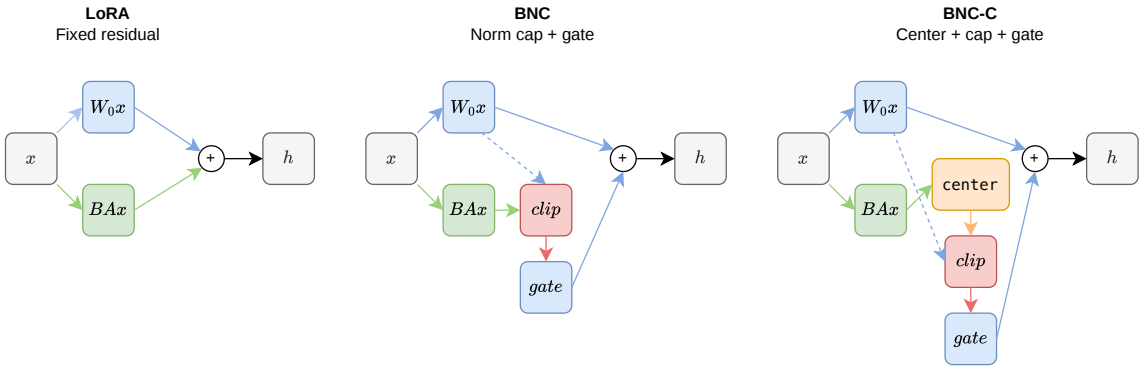


Figure 1: Implementation-faithful comparison of LoRA, BNC, and BNC-C. BNC clips the adapter update against the frozen base-output norm and then applies a bounded learnable gate; BNC-C inserts per-token mean-direction centering before the same clip-and-gate path.

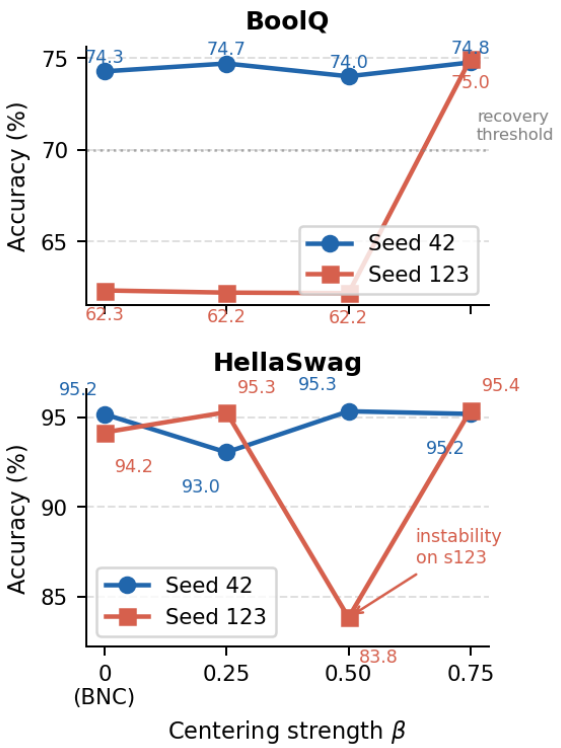


Figure 2: Centering strength β ablation at rank 8 (BoolQ top, HellaSwag bottom). $\beta = 0$ is pure BNC; BoolQ seed 123 recovers only at $\beta = 0.75$.

Fine-tuning. Commonsense 170K dataset (Hu et al., 2023); 3 epochs, learning rate 1×10^{-4}, batch size 16, target modules `[q,k,v,up,down]_proj`, $\alpha = 2r$. All methods use the same hyperparameters and three random seeds (42, 123, 2024).

External baselines. LoRA, DoRA (Liu et al., 2024), and RsLoRA (Kalajdzievski, 2023), all using the same training setup.

Author comparator. We additionally include DyS-GS-Res only as an internal comparator against an earlier coupled-gate design.

Evaluation. Zero-shot accuracy on eight benchmarks: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), Social IQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC-Challenge and ARC-Easy (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018).

4.2 Two Collapse Types: A Motivating Observation

Figure 3 shows per-seed accuracy of standard LoRA at rank 8. Seed 123 is the weakest run overall, but the most diagnostically sharp failures are HellaSwag and BoolQ: HellaSwag collapses to 54.7% (near chance for a 4-way task) and BoolQ drops to 62.8% (near the majority baseline of 61.9%). Other tasks also vary by seed, but their degradation is milder and does not exhibit the same chance-like or majority-collapse signature.

Collapsed and stable seeds are also difficult to distinguish under standard optimisation metrics: in the inspected runs, the collapse only becomes clear after held-out evaluation (Appx. Fig. 6).

Table 1 shows the critical diagnostic on the fragile seed 123 case. Default BNC and DoRA both resolve HellaSwag collapse (54.7% $\rightarrow$ 94.2% and 95.4% respectively) while leaving BoolQ near the majority baseline. The BoolQ rows sharpen the second regime: the default rank-8 methods collapse, but the same seed can be recovered by tightening the local BNC constraint ($\nu = 0.5$), switching to TACL, a partial TAC-style geometry-control preset, or to full TAC, or applying soft centering (BNC-C,

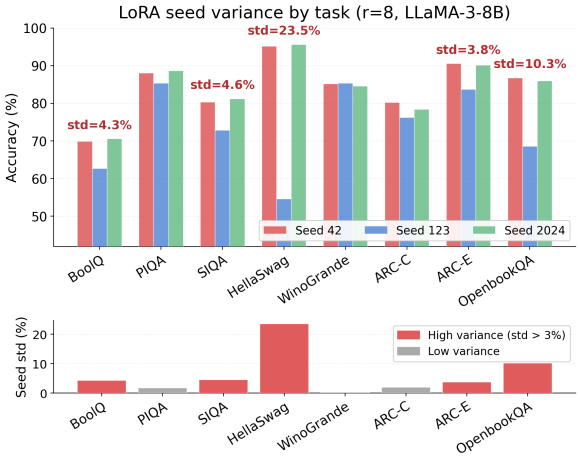


Figure 3: LoRA per-seed accuracy and cross-seed std on eight commonsense tasks (r=8, LLaMA 3-8B, seeds 42/123/2024). Seed 123 shows the sharpest failures on HellaSwag and BoolQ.

Method	BoolQ	Hella.
LLaMA 3-8B (no FT)	60.70%	25.00%
LoRA	62.8%	54.7%
DoRA	62.2%	95.4%
BNC	62.3%	94.2%
BNC ($\nu=0.5$)	73.9%	95.1%
TACL	66.8%	95.5%
TAC	74.6%	95.5%
BNC-C ($\beta=0.75$)	**74.9%**	95.4%

Table 1: BoolQ and HellaSwag accuracy at seed 123, rank 8. HellaSwag is repaired by most methods; BoolQ requires tighter BNC ($\nu = 0.5$), TAC, or centering (BNC-C).

$\beta=0.75$). Table 2 further decomposes this fragile seed 123 case into same-rank mechanism probes, isolating which geometry-controlled variants remain on the shared partial-recovery plateau.

4.3 BNC Results: Fixing Amplitude-Driven Collapse

Table 3 reports multi-seed results at rank 8. BNC achieves $84.4\% \pm 1.0\%$ average accuracy across three seeds. The elevated std is dominated by BoolQ at seed 123; for the seven non-BoolQ tasks the cross-seed std stays below 1.7 points.

Most strikingly, HellaSwag is stabilised across the three canonical seeds: BNC stays near 95% on all three seeds, compared to LoRA's 54.7% collapse at seed 123. This demonstrates that the observed amplitude-driven collapse can be prevented by per-token norm calibration.

BoolQ remains the exception for the default BNC setting ($\nu=1.0$): BNC achieves 74.3% (s42)

Method	BoolQ	Hella.
TR	66.8%	95.5%
UGC-Center-only	66.8%	95.5%
UGC-Shape-only	66.8%	95.5%
TACL	66.8%	95.5%
TAC	74.6%	95.5%
BNC ($\nu=0.5$)	73.9%	95.1%

Table 2: Same-rank ($r=8$) mechanism probes on the fragile BoolQ seed 123 case. TR, UGC-Center-only, UGC-Shape-only, and TACL share the same partial-recovery plateau, while only full TAC exits it within the UGC family.

and 73.6% (s2024), but only 62.3% at seed 123—essentially the same as LoRA on the same seed. Figure 4 sharpens the diagnosis: BNC on seed 123 predicts "true" on 99.6% of BoolQ examples, collapsing to the majority baseline despite high confidence. This pattern is consistent with a mean-direction bias in the LoRA update: the per-token row mean dominates the update and pushes predictions toward one answer.

BNC-C ($\beta = 0.75$) corrects this directly. By subtracting 75% of the per-token mean direction before the norm clip, it reduces the observed directional bias without destabilising the amplitude constraint. BoolQ seed 123 recovers to 74.9%, while HellaSwag remains at 95.4% across all seeds. Overall, BNC-C achieves $84.9\% \pm 0.1\%$ average accuracy—the BoolQ cross-seed std drops from 6.8% to 0.7%, and the overall std from 1.0% to 0.1%. Section 3.2 describes the method and the β ablation in detail.

4.4 Rank-Threshold Analysis: A Boundary-Approximate BoolQ Regime

The BoolQ seed 123 collapse at rank 8 is shared by standard LoRA-style baselines and direct norm-control methods: LoRA, DoRA, and BNC all remain near the majority baseline on this setting. This shared failure suggests a rank-sensitive regime rather than a simple method-specific one.

Table 4 tests the rank-threshold hypothesis by reporting BoolQ at seed 123 for $r \in \{8, 16\}$. Every completed baseline run that we have evaluated at rank 16 recovers above 73%—a 10+ point improvement from the same rank-8 collapsed state. BNC also recovers at rank 16 (75.0%), which supports a real rank-threshold effect for this method family. At the same time, recent TAC results show that the same rank-8 setting can be recovered without increasing rank, and tighter-BNC runs show that

Method	BoolQ	PIQA	SIQA	Hella.	Wino.	ARC-c	ARC-e	OBQA	Avg
No fine-tuning									
LLaMA 3-8B	60.70	41.22	28.33	25.00	10.86	15.98	16.50	16.60	26.90
External baselines (3 seeds)									
LoRA	67.7±4.3	87.4±1.8	78.1±4.6	81.8±23.5	85.1±0.4	78.3±2.0	88.2±3.8	80.5±10.3	80.9±6.2
rsLoRA	69.7±6.5	88.3±0.3	81.2±0.5	95.1±0.1	85.0±1.0	79.2±0.9	89.3±1.3	83.7±3.1	83.9±0.7
DoRA	66.6±7.4	88.4±0.2	80.2±0.9	95.4±0.0	85.1±0.6	79.5±1.1	90.4±0.1	85.5±1.3	83.9±1.0
Ours (3 seeds)									
DyS-GS-Res	68.7±6.0	85.1±5.7	78.2±5.1	80.6±17.0	83.8±2.4	71.8±13.6	83.2±11.9	80.1±9.6	78.9±8.8
BNC	70.1±6.8	88.2±0.2	80.9±0.4	95.0±0.8	84.5±1.3	**79.9**±0.5	**90.0**±0.4	**86.5**±1.7	84.4±1.0
BNC-C	**74.5**±0.7	**88.4**±0.5	**81.1**±0.1	**95.4**±0.2	**85.3**±0.4	79.3±0.5	89.9±0.2	85.3±0.8	**84.9**±0.1

Table 3: Mean ± std accuracy (%) over three canonical seeds (42, 123, 2024) at rank 8 on LLaMA 3-8B. BNC-C uses $\beta = 0.75$. Bold marks the best author-method result in each column.

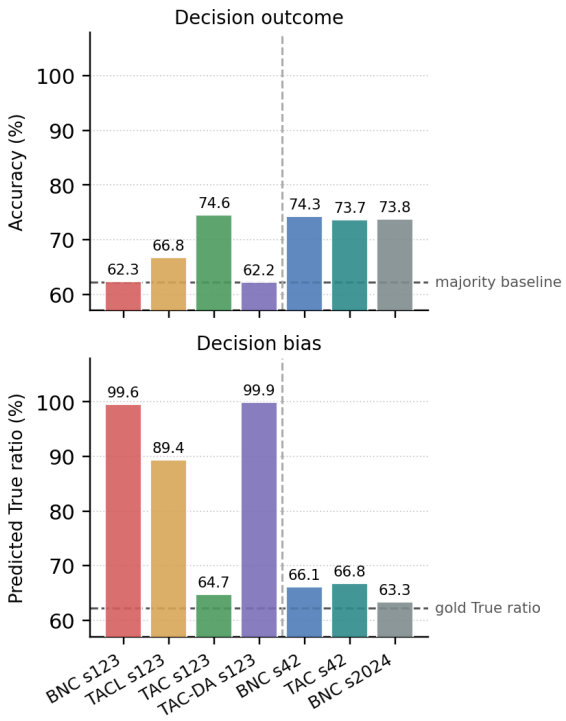


Figure 4: BoolQ diagnosis at rank 8 (seed 123 variants left; stable seeds right). Top: accuracy; Bottom: predicted-"true" ratio.

Method	$r=8$ (s123)	$r=16$ (s123)
LoRA	62.8%	73.8%
DoRA	62.2%	75.2%
DyS-GS-Res	62.2%	74.9%
BNC	62.3%	75.0%
BNC-C	**74.9%**	74.3%‡

Table 4: BoolQ accuracy at seed 123 for $r \in \{8, 16\}$. Standard methods collapse at $r=8$ but recover at $r=16$; BNC-C recovers at both ranks without requiring a rank increase. ‡$r=16$ uses $\beta=0.5$ (see §4.5).

stronger local constraints can also recover the same case at fixed rank. The phenomenon is therefore better viewed as a soft rank-sensitive regime than as a hard capacity wall.

Table 4 reports BoolQ accuracy for seed 123 only, isolating the critical fragile case; the corresponding 8-task means are reported in Table 5. The $r = 16$ calibrated BNC-C entry (74.3%) is not exceptional on BoolQ alone—its value is that it removes the failure while restoring stable eight-task performance ($85.06\% \pm 0.04\%$, Table 5). We interpret this as follows. At rank 8, the adapter column space may be insufficient to represent the decision boundary for the gradient trajectory seeded by seed 123. The completed rank-16 runs suggest that additional dimensions are sufficient to represent the boundary for this setting. However, the same low-rank budget can also succeed under full TAC, while simpler UGC variants remain on the same lower plateau for this seed (Table 2); tighter-BNC ablations suggest that stronger local constraints can also be sufficient. The bottleneck is therefore not simply "more rank or fail", and for binary classification tasks like BoolQ, rank selection and update structure may matter more than the specific PEFT label. If multiple methods collapse on the same task-seed combination, increasing rank is a useful diagnostic; but if a structured low-rank update recovers the same case, the failure should not be described as a hard capacity wall.

4.5 Main Results Across Ranks

Table 5 shows average accuracy at $r \in \{8, 16, 32\}$. BNC is the most consistently stable method across all three ranks: its mean improves monotonically from $84.4\% \pm 1.0\%$ at $r = 8$ to $84.8\% \pm 0.7\%$ at $r = 16$ and $85.2\% \pm 0.1\%$ at $r = 32$, with cross-seed variance shrinking at every step. In contrast,

LoRA shows large-variance collapses at both $r=8$ ($\pm 6.2\%$) and $r=32$ ($\pm 3.9\%$), and rsLoRA degrades steadily as rank increases. DoRA achieves the best single entry at $r=16$ ($85.5\% \pm 0.2\%$), outperforming BNC at that rank by 0.7%; at $r=8$ and $r=32$, BNC is stronger ($+0.5\%$ and $+0.6\%$ respectively). The key advantage of BNC is therefore not a single-rank peak but its absence of catastrophic collapses across the full rank range tested.

Method	$r=8$	$r=16$	$r=32$
External baselines			
LoRA	80.89±6.21	85.03±0.13	82.72±3.90
rsLoRA	83.93±0.75	83.81±0.21	81.73±0.48
DoRA	83.88±1.05	**85.50**±0.24	84.57±0.64
Ours			
BNC	84.40±1.02	84.82±0.66	85.15±0.13
BNC-C	**84.90**±0.12	85.06±0.04‡	**85.52**±0.30

Table 5: Average accuracy mean ± std (%) over three canonical seeds (42, 123, 2024) for $r \in \{8, 16, 32\}$. Bold marks the highest value in each column. † marks a setting exhibiting training-dynamics instability (eval-time β override has zero effect).

Note: LoRA at $r=32$ (seed 42) and $r=8$ (seed 123) each show a single-seed collapse driving the large variance. ‡BNC-C at $r=16$ uses $\beta=0.5$; the stronger setting $\beta=0.75$ over-corrects at this rank ($83.6\% \pm 1.4\%$, training-induced). See §6 for details.

The centering extension BNC-C yields a rank-dependent picture whose key variable is the training-time centering strength β. At $r=8$ and $r=32$, the stronger setting $\beta=0.75$ outperforms BNC ($84.9\% \pm 0.1\%$ at $r=8$; $85.5\% \pm 0.3\%$ at $r=32$, the highest result across all methods and ranks we test). At $r=16$, however, $\beta=0.75$ over-corrects: accuracy drops to $83.6\% \pm 1.4\%$, below every baseline. Reducing centering strength to $\beta=0.5$ fully restores the $r=16$ regime: all three seeds recover to 85.1%, 85.1%, 85.0%, yielding $85.06\% \pm 0.04\%$—above BNC at $r=16$ (84.8%) and near-zero cross-seed variance. That eval-time β overrides ($\beta \in \{0, 0.25, 0.5, 0.75\}$) left the failed $\beta=0.75$ checkpoints unchanged confirms that centering acts as a training-time update-shaping effect rather than an inference-time correction.

In summary, across the three ranks we test, the best-performing calibrated operating point of BNC-C outperforms BNC in every case. We note that this comparison is internal to the BNC family: baselines were not subjected to equivalent per-rank retuning, so the comparison reflects BNC-C's rank-sensitive operating point rather than a fully matched calibration study. BNC remains the recommended zero-configuration default: it improves on LoRA and rsLoRA at every rank without rank-specific failure modes and requires no hyperparameter search beyond the standard LoRA setup.

4.6 Training Efficiency

Table 6 compares training time and trainable parameter count. BNC adds 160 scalar gate parameters ($< 0.001\%$ of total) and negligible forward-pass overhead (per-token norm computation is dominated by the matrix multiplications). DoRA adds $1.4\times$ overhead due to per-step column-norm recomputation.

Method	Params	Avg train time	vs LoRA
LoRA	56,623,104	11h 18m	1.00×
BNC	56,623,264	≈11h 18m	≈1.00×
DoRA	57,409,536	16h 08m	1.43×

Table 6: Training efficiency at rank 32 (LLaMA 3-8B). LoRA and DoRA times are averaged over 3 seeds; BNC is an implementation-level runtime estimate.

5 Related Work

Parameter-Efficient Fine-Tuning. Beyond LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024), PEFT methods include Adapter layers (Houlsby et al., 2019), Prefix Tuning (Li and Liang, 2021), and $(IA)^3$ (Liu et al., 2022). BNC instead constrains the LoRA update relative to the base output.

Rank-Stabilised and Adaptive Scaling. rsLoRA (Kalajdzievski, 2023) provides a better fixed scaling scalar across ranks. AdaLoRA (Zhang et al., 2023) and SoRA (Ding et al., 2023) focus on adaptive rank allocation, orthogonal to per-token norm calibration. Our rank-threshold analysis complements adaptive-rank methods by empirically identifying task-specific rank requirements via cross-seed comparison. Label smoothing (Müller et al., 2019) regularises overconfident predictions and can improve fine-tuning stability.

6 Discussion

Diagnostic implications. An engineer who observes collapsed BoolQ accuracy at $r=8$ might attempt method switching (LoRA → DoRA → default BNC) only to find no improvement, because the failure is near a soft rank limit. Our rank-threshold analysis provides a more efficient

diagnostic: if all methods collapse on the same task-seed combination, first test BNC-C with calibrated β or a modest rank increase before discarding the entire method family. When rank-specific β tuning is feasible, the calibrated BNC-C operating point is the best-performing variant in our experiments; BNC is the zero-configuration default otherwise.

Mechanistic evidence from sim-to-base analysis. Although the two collapse types present differently at the label level, a cross-dataset similarity analysis reveals a common low-level signature. We compute the cosine similarity between each model's predicted label distribution and the untuned base model's distribution (*sim-to-base*) for each of the eight evaluation tasks. Healthy fine-tuned models achieve *sim-to-base* ≈ 0.49–0.50 on HellaSwag, indicating the learned representations have moved substantially away from the pretrained prior. By contrast, two representative collapsed runs are shown in Appx. Table 7: a fully collapsed early TAC run reaches *sim-to-base* $= 1.00$ on *all* eight datasets simultaneously, and a partially collapsed LoRA run (seed 123, $r = 8$) reaches 0.91 on HellaSwag while still achieving near-chance accuracy (55.8%). On HellaSwag specifically, these numbers suggest that severe collapse reflects *fine-tuning erasure*—the base model's pretrained prediction prior re-emerges as activation amplitudes grow—rather than a training artifact that happens to favour one label. We offer this as supporting mechanistic evidence for the amplitude-blowup interpretation; partial collapses and rank-sensitive failures indicate a more complex failure space than a single axis.

Practical implications. DoRA remains a strong baseline and achieves the best single-rank mean at $r = 16$ (85.5%), but BNC offers a cleaner robustness-efficiency trade-off: it matches standard LoRA in trainable parameter count and runtime (Table 6), attains much lower variance at $r = 32$ ($\pm 0.13\%$ vs. $\pm 0.64\%$), and shares with DoRA the same unresolved rank-8 BoolQ failure that motivates the additional centering step. In the representative runs we inspected, collapse is not reliably signalled by standard optimisation metrics: the collapsed seed can achieve validation loss comparable to or lower than stable seeds, so loss alone does not predict downstream recovery (Appx. Fig. 7). Multi-seed downstream evaluation is therefore essential for both collapse types.

Rank-sensitivity of centering. Centering is a rank-sensitive training-time intervention rather than a monotonic improvement over norm calibration. Eval-time β overrides on the failed $r = 16$, $\beta = 0.75$ checkpoints produce no change, confirming that the effect is baked into the training trajectory rather than acting as a post-hoc inference correction. To test the baseline side of this fragile regime, we trained LoRA with label smoothing ($\varepsilon = 0.1$) on the critical $r = 8$, seed 123 configuration; BoolQ and HellaSwag accuracy were unchanged (62.8% and 54.7% respectively). This suggests that the low-rank collapse regime is sensitive to training-time configuration, and that the minimal output-regularisation repair we tested does not provide the right inductive bias, while training-time centering can.

7 Conclusion

We studied seed-dependent collapse in LoRA fine-tuning on LLaMA 3-8B commonsense reasoning and found that a single PEFT intervention does not explain all observed failures. BNC isolates the amplitude-sensitive regime through per-token norm calibration, restoring HellaSwag seed 123 from 54.7% to 94.2%, while BNC-C extends it with soft centering that recovers BoolQ seed 123 from 62.3% to 74.9% in the boundary-sensitive regime.

Seed-dependent accuracy collapse should not automatically be attributed to the fine-tuning method; a soft rank threshold may be the true bottleneck. Our collapse taxonomy provides a systematic diagnostic: if multiple methods fail on the same task-seed combination, first test stronger local constraints, soft centering, or a modest rank increase before assuming the entire method family is broken.

Limitations

Our experiments are limited to LLaMA 3-8B and commonsense reasoning. Although we identify stable operating points for BNC-C at each tested rank, $\beta = 0.75$ over-corrects at $r = 16$ while $\beta = 0.5$ behaves differently at $r = 8$, so the current method requires per-rank calibration rather than a universal β or an automatic schedule. The precise mechanism by which seed 123 triggers the BoolQ boundary-approximate collapse at $r = 8$, and what determines the rank-dependent sensitivity of β, remain open questions for future work.

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of AAAI*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. SoRA: Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. LLM-Adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Damian Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*.

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP*.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQA: Commonsense reasoning about social interactions. In *Proceedings of EMNLP*.

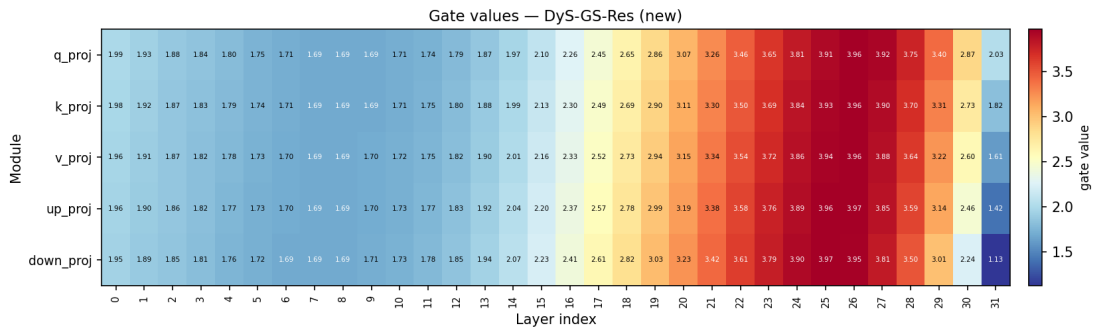
Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*.

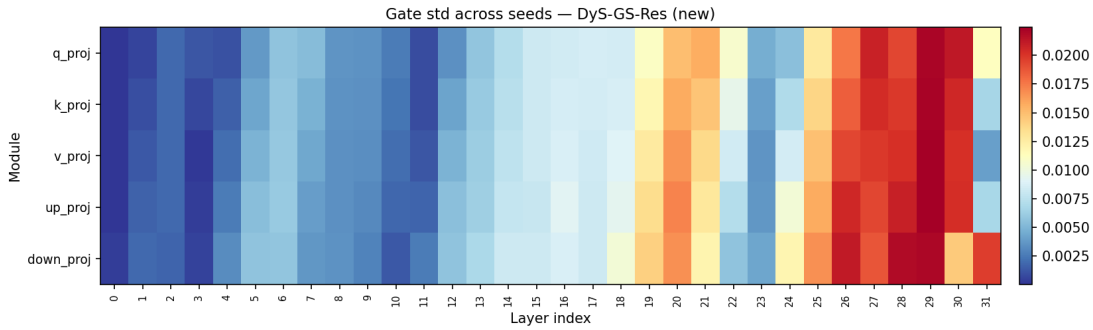
A Supplementary Visualisations

Model	Collapse	*sim-to-base*	HellaSwag acc
Base model (no FT)	—	1.00	25.0%
TAC (complete)	full	1.00	25.0%
LoRA r8 s123	partial	0.91	55.8%
LoRA r8 s42	none	0.49	95.2%
BNC s2024	none	0.50	95.6%

Table 7: Supplementary cosine similarity between each model's HellaSwag prediction distribution and the untuned base model (*sim-to-base*). Complete collapse recovers the base prior exactly; partial collapse sits between base and healthy fine-tuning.



(a) Gate values (r=32, seed=42).



(b) Cross-seed std (r=16, 3 seeds).

Figure 5: Supplementary DyS-GS-Res gate visualisation: depth increases the learned gates, while cross-seed std stays below 0.02.

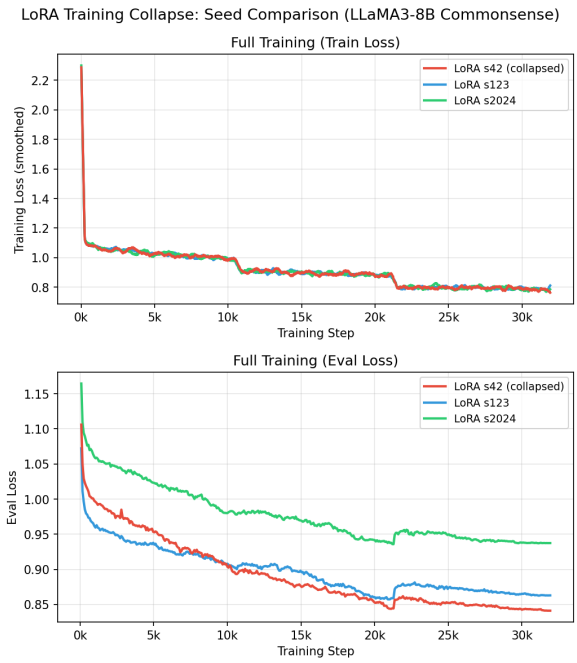


Figure 6: Supplementary training dynamics from an analogous low-rank LoRA run (r=32, LLaMA 3-8B, three seeds): collapsed and stable seeds show similar losses.

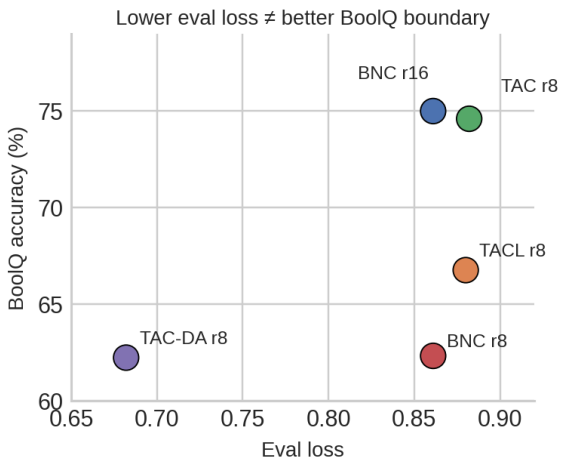


Figure 7: Supplementary eval loss vs. BoolQ accuracy for same-rank interventions (seed 123, r=8). Lower eval loss does not predict BoolQ recovery.

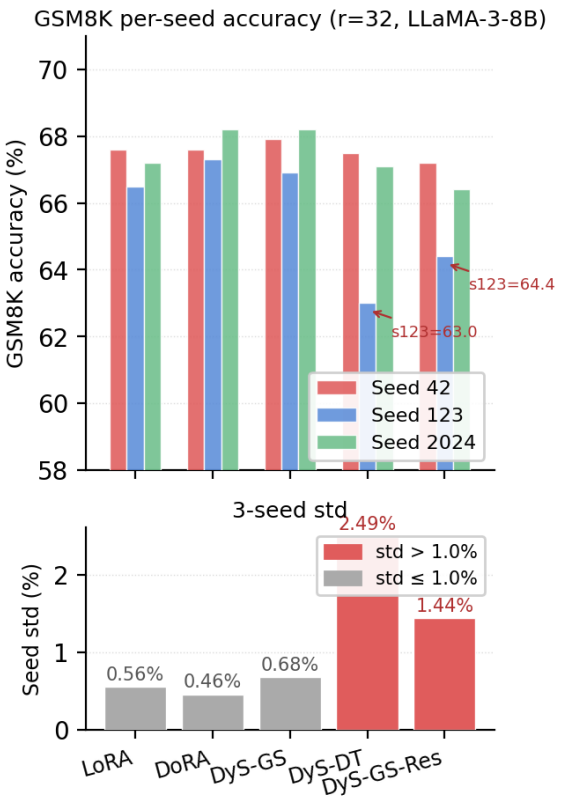


Figure 8: Supplementary GSM8K per-seed accuracy (top) and three-seed standard deviation (bottom) at r=32 on LLaMA 3-8B. This side observation is not part of the main collapse taxonomy claim.