NEURIPS 2026 · MAIN CONFERENCE

Beyond Parameter Arithmetic

Sparse Complementary Fusion for Distribution-Aware Model Merging

Weihong Lin[1] · Qilong Shi[3] · Aomufei Yuan[2] · Yuxuan Tian[2] · Zhengyang Wang[2]
Guangxiang Zhao[1] · Xiangzheng Zhang[1] · Tong Yang[2] · Lin Sun[1,*]

[1]Beijing Qiyuan Technology [2]Peking University [3]Tsinghua University · *sunlin1@360.cn



Merging is training-free — until the model can’t stop.

97–100%

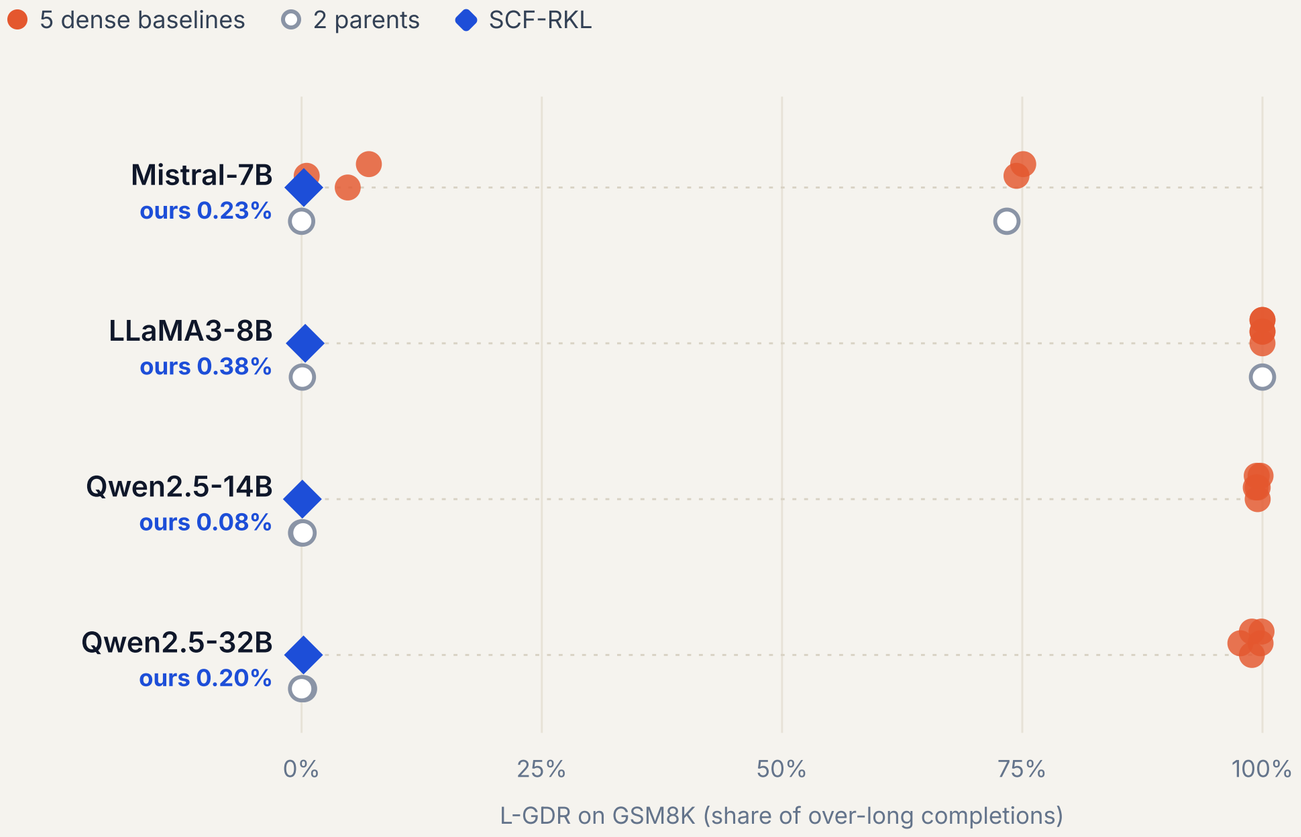
of GSM8K completions are abnormally long for **five dense baselines** at 14B / 32B — although both parents are clean (≤ 0.2%).

−70 pts

accuracy lost on 32B GSM8K. Over-long outputs also exhaust the context in multi-turn use.

< 1%

SCF-RKL on all four pairs, even when one parent is already 73–100% dysregulated.



Paper Fig. 1 (left) redrawn · L-GDR@5× under the reference protocol.

The baseline often knows the answer. It just never stops.

CASE 1 · STRUCTURAL REPETITION

Wendi feeds each chicken 3 cups a day in three meals: 15 cups in the morning, 25 in the afternoon. How many cups in the final meal for 20 chickens?

SCF-RKL ✓

20 × 3 = 60 cups.
15 + 25 = 40 cups.
60 − 40 = 20 cups.
Answer: 20 cups.

⏎ stops normally (EOS)

DARE-LINEAR ✗

60 − 40 = 20 cups.

```
**Final Answer:** \boxed{20}
**Final Answer:** \boxed{20}
**Final Answer:** \boxed{20}
**Final Answer:** \boxed{20}
**Final Answer:** \boxed{20}
**Final Answer:** \boxed{20}
...
```

∞ repeats until max tokens

CASE 2 · SEMANTIC COLLAPSE

Which is sarcastic? (A) “He’s been *drinking* every day recently, what a great transformation.” (B) “He’s been *exercising* every day recently, what a great transformation.”

SCF-RKL ✓

(A) calls daily drinking a “great transformation” — ironic. (B) sincerely praises exercise.
Answer: (A)

⏎ stops normally (EOS)

DARE-LINEAR ✗

“Let’s think step by step. If we look at (A), it says that he’s been drinking…”

```
Wentw to examples totheG
inthew evident 1 DandC A D …
theD D to theD:D. s D D D D M
M wP D D hD orFthat t …
```

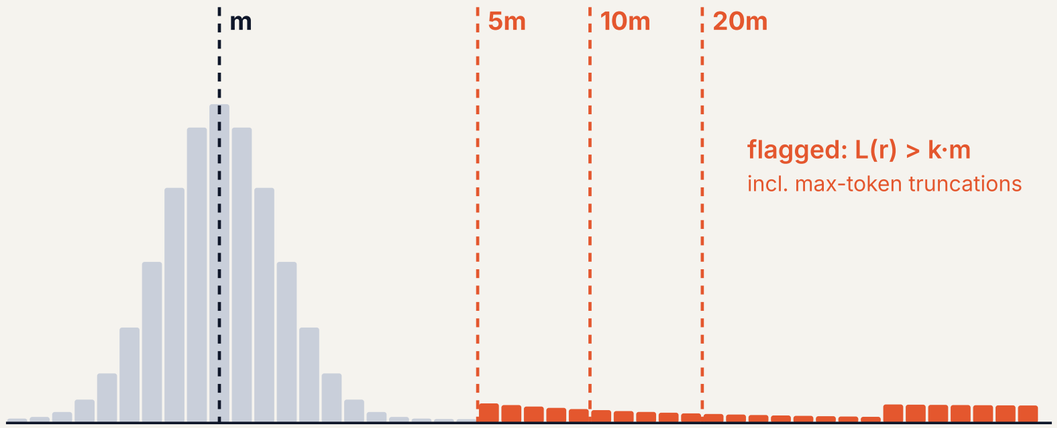
↯ drifts into symbols

SPECTRUM structural repetition · termination failure · circular reasoning · semantic collapse

A simple length screen: L-GDR

Flag completions that are far longer than what the **unmerged parents** produce on the same benchmark.

$$\text{L-GDR@}k\times = \frac{1}{N}\sum_{r=1}^{N} 1\{\, L(r) > k\,m \,\}$$



ROBUST TO DECODING SETTINGS · MISTRAL-7B, GSM8K

TEMP.	@5×	@10×	@20×
0.0	**1.36** / 6.60	**1.29** / 6.37	**1.29** / 6.29
0.6	**1.76** / 14.52	**0.99** / 13.32	**0.59** / 12.36
1.0	**2.96** / 16.68	**0.61** / 9.63	**0.15** / 4.70

Cells: **SCF-RKL** / TSV-Merge L-GDR (%), matched runs.

$k \in \{5, 10, 20\}$; m: median completion length of the parents (T = 0.6, n = 4, max_new_tokens = 4096). Always read next to accuracy, max-hit, normal-stop, Rep-4 and a semantic audit — it is a screen, not a label.

All nine paired L-GDR-gap confidence intervals favour SCF-RKL, across temperature and threshold.

Averaging **invents** parameters.

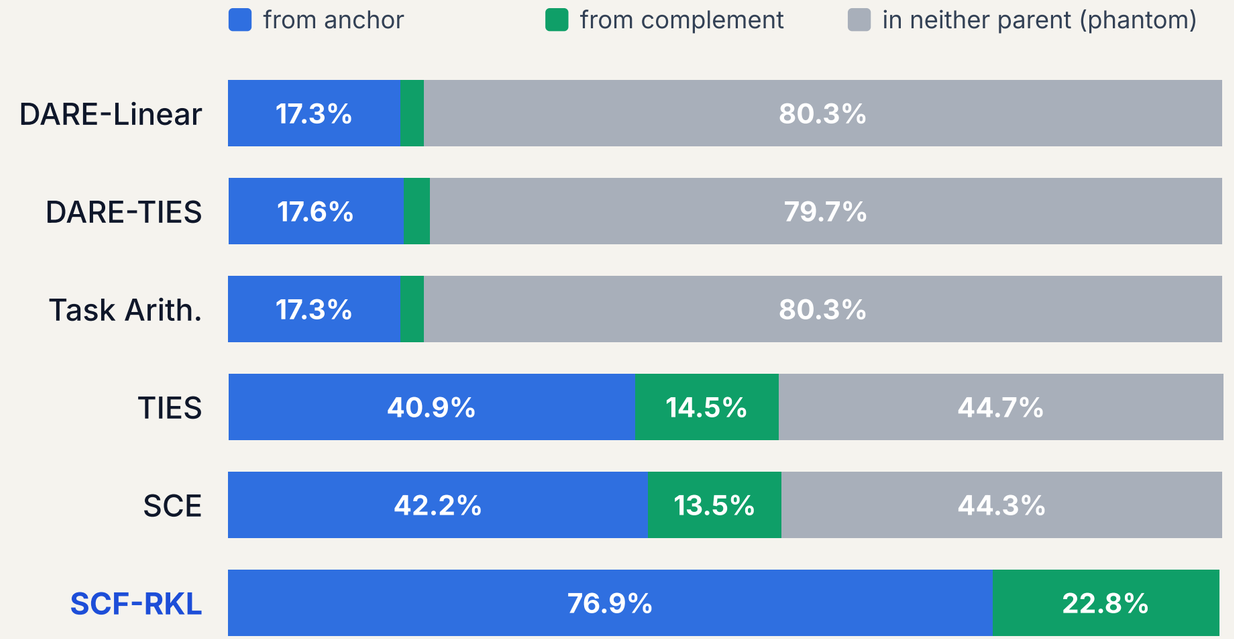
Any convex mix $\lambda\theta_a + (1-\lambda)\theta_c$ creates values that exist in **neither parent** — on every coordinate where they differ.

0.82	-0.31	0.26
anchor θ_a	complement θ_c	average · phantom

Dense merging perturbs almost every coordinate at once, with values the network never saw.

Selection keeps 0.82 *or* copies −0.31. Every fused value comes from a parent (Prop. 3.3).

WHERE FUSED VALUES COME FROM · LAYER 15 (PAPER FIG. 2D)



Dense methods: **44–80%** of values in neither parent. SCF-RKL: values from the two parents only.

Three rules: select, sparsely, anchor first.

1

Binary selection, not interpolation

Copy each coordinate from exactly one parent. No phantom values by construction.

2

Low-density updates

Replace only a statistical tail, so entropy drift and subspace rotation stay bounded.

3

Anchor-centric saliency

Spend the budget where the parents' row-wise proxies really differ, measured from the anchor's side.

0%

fused values outside the parents (dense: 44–80%)

5.43%

of coordinates replaced (Mistral-7B, α = 1.5; 5.05–6.60% per layer)

0.5% vs 72.3%

L-GDR of RKL selection vs. a random binary mask

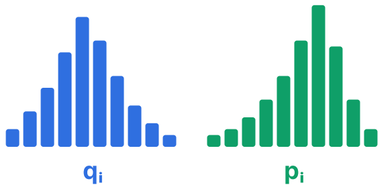
SCF-RKL in four steps, tensor by tensor.

STEP 1

Row-wise proxy

$$q_i = \mathrm{softmax}(\theta_{a,i,:})$$

$$p_i = \mathrm{softmax}(\theta_{c,i,:})$$

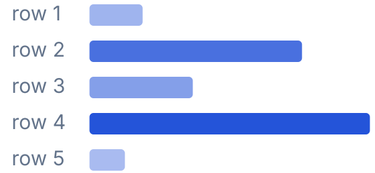


Each output row becomes a distribution — a weight-space proxy, not a token distribution.

STEP 2

Reverse KL per row

$$D_{\mathrm{KL}}^{\mathrm{rev}}(q_i \| p_i) = \sum_j q_{i,j} \log \frac{q_{i,j}}{p_{i,j}}$$

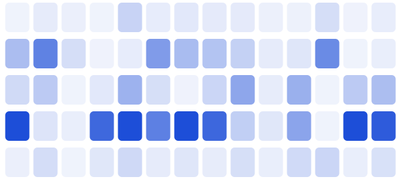


Weighted by the **anchor** proxy: rows that truly shift get more weight.

STEP 3

Saliency

$$\mathcal{I}_{i,j} = |\Delta_{i,j}| \cdot D_{\mathrm{KL}}^{\mathrm{rev}}(q_i \| p_i)$$



$|\Delta|$ ranks coordinates within a row; RKL re-weights across rows (Fisher-inspired).

STEP 4

IQR mask & fuse

$$\tau = \mathrm{med}(\mathcal{I}) + \alpha\,\mathrm{IQR}(\mathcal{I})$$

$$\theta_f = \theta_a + 1[\mathcal{I} \geq \tau] \odot \Delta$$



α = 1.5 everywhere. No calibration data, no gradients.

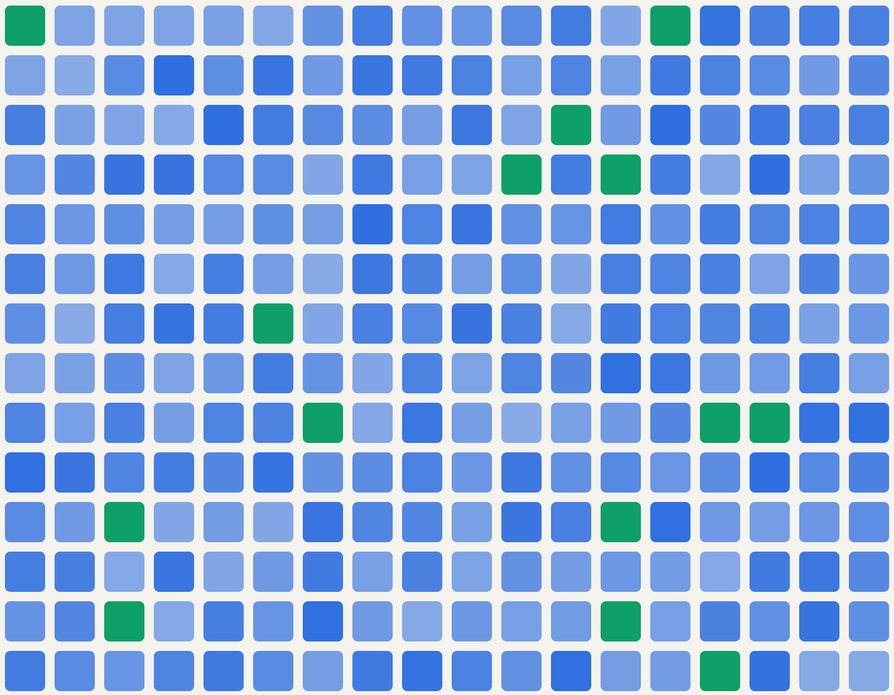
```
# per weight tensor W (anchor θa, complement θc)
Δ   = |θc - θa|
q, p = softmax(θa, dim=-1) + ε, softmax(θc, dim=-1) + ε
rkl = (q * log(q / p)).sum(-1, keepdim=True)   # one scalar per row
I   = Δ * rkl
τ   = median(I) + 1.5 * (Q3(I) - Q1(I))
θf  = θa + (I ≥ τ) * (θc - θa)              # binary: no new values
```

parameter-only · one-shot · offline · O(d log d)

H800	TIME	PEAK GPU
7B	59 s	4.2 GB
14B	331 s	24.7 GB
32B	352 s	24.7 GB

≈ 1.6–2.2× the wall-clock of Task Arithmetic, paid once at merge time.

Sparse, binary — and bounded.



kept from anchor (94.6%) copied from complement (5.4%)

Realised density on the audited Mistral-7B pair; it varies with tensor and model pair.

PROP. 3.3 · NO PHANTOM PARAMETERS

$$\theta_{f,i} \in \{\theta_{a,i}, \theta_{c,i}\} \;\; \forall i$$

Interpolation with $\lambda \in (0,1)$ creates new values wherever the parents differ.

PROP. 3.1 · ENTROPY PRESERVATION

$$H(q_{f,i}) \;\geq\; H(q_i) - L \sum_{j \in S_i} |\Delta_{i,j}|$$

Masking can only shrink each row's ℓ_1 perturbation.

THM. 3.2 · BOUNDED SUBSPACE ROTATION

$$\sin \Theta_{\max}(U, \tilde{U}) \;\leq\; \frac{\sqrt{2}\,\|E\|_2}{\delta - \|E\|_2}, \quad E = M \odot (W_c - W_a)$$

Mistral-7B at α = 1.5: $\|E\|_2/\delta < 0.3$. Measured rotation stays **< 40°** vs. ≈ 80° for dense methods; NSS ≈ 10^{-3}.

These are proxy-space bounds that hold for any mask — motivation for the design, not an end-to-end generation guarantee.

Four fusion scenarios. One fixed setting.

A · SIMILAR CAPABILITY

14B / 32B

DeepSeek-R1-Distill-Qwen2.5 → TinyR1 experts: Math + Code (14B), Math + Science (32B)

B · STRONG-WEAK

7B / 8B

Mistral-7B-Instruct + MathCoder2; Llama-3-8B-Instruct + MAmmoTH2-8B-Plus (gap ≈ 15 pts)

C · REASONING + SAFETY

32B

A safety-hardened model fused with a reasoning-specialised one; 13 benchmarks

D · VISION

8-in-1 ViT

ViT-B/32: four 2-in-1, two 4-in-1, one final merge

9 PARAMETER-ONLY BASELINES

Task Arithmetic · DARE-Linear · TIES · DARE-TIES · SCE · TSV-Merge · ISO-CTS · DELLA · Model Breadcrumbs · 7B/8B pairs

AdaMerging and LED-Merging need task data, so they are out of scope.

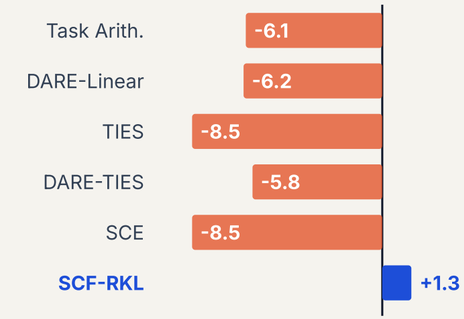
23 BENCHMARKS

AIME24 / 25 · GPQA · LiveCodeBench v5 · GSM8K · BBH · MMLU · HumanEval · IFEval · IFBench · 5 safety suites · 8 vision tasks

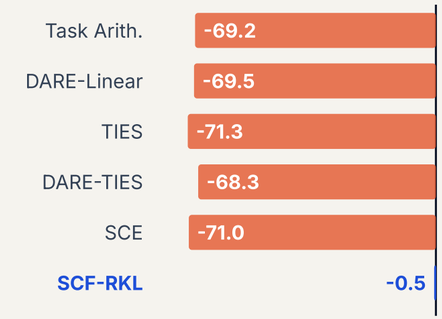
$\alpha = 1.5$, $\varepsilon = 10^{-8}$ for every pair — no per-pair tuning.

The only positive average at 14B and 32B.

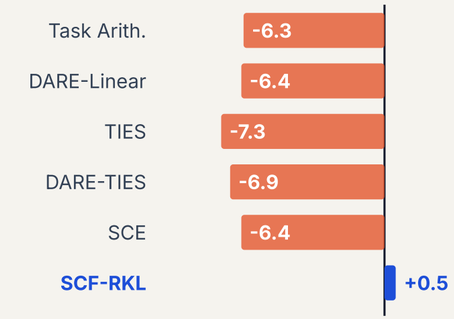
32B · 10-TASK AVG CHANGE (PP)



32B · GSM8K CHANGE (PP)



14B · 10-TASK AVG CHANGE (PP)

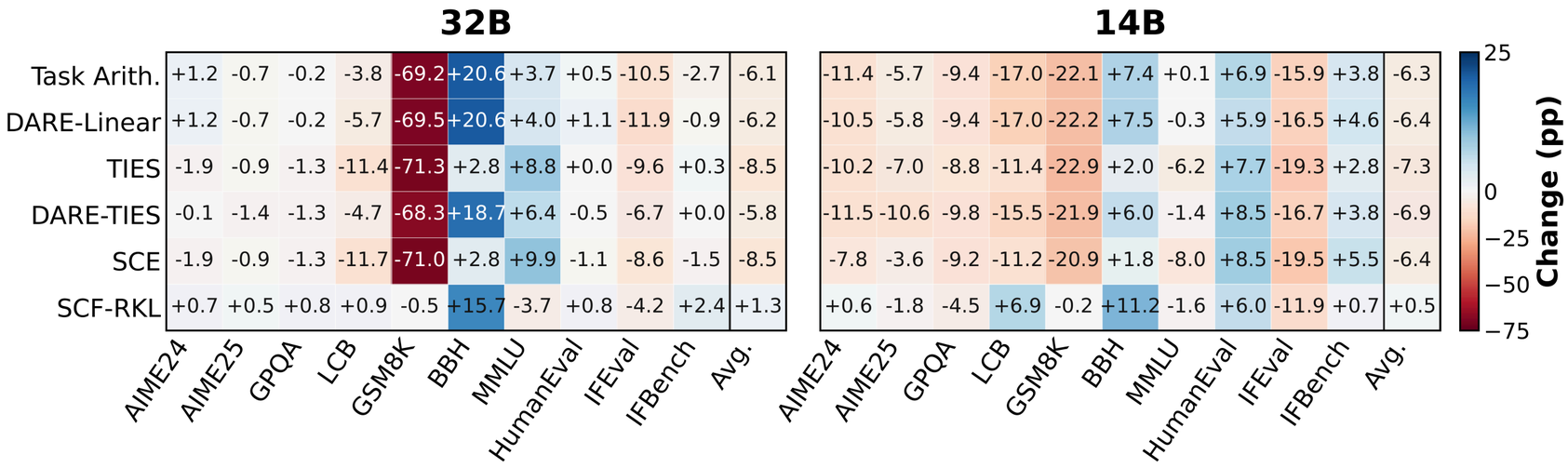


7 / 10

32B benchmarks improved (baselines ≤ 4)

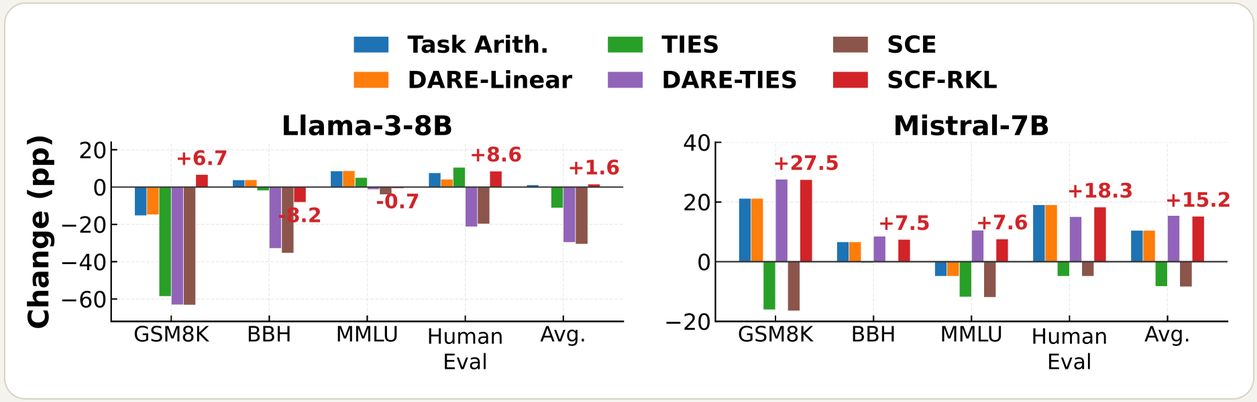
+6.55

14B LiveCodeBench over the better parent (62.46 vs 55.91)



Stable even with a broken parent.

CHANGE FROM THE TWO-PARENT MEAN (PP) · PAPER FIG. 8



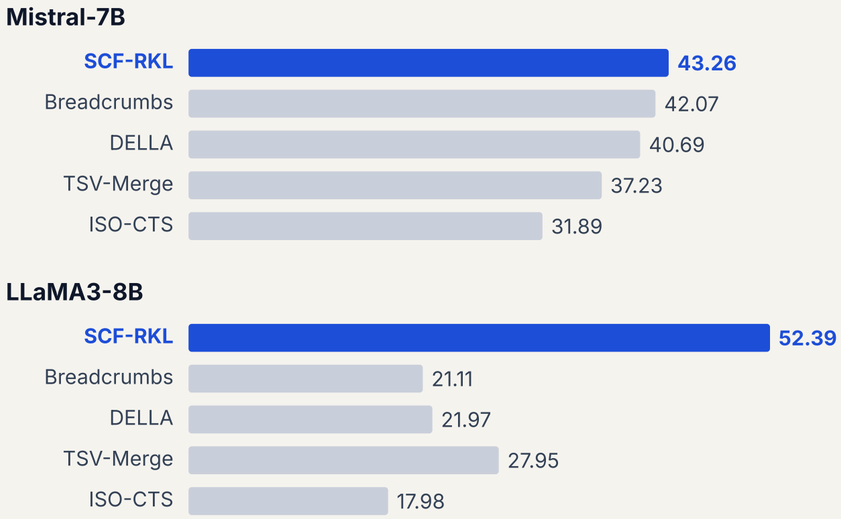
+15.2

Mistral-7B: above the parent mean on all four tasks.

0.38%

LLaMA3-8B: baselines inherit MAmmoTH2's 100% L-GDR; we gain +6.7 GSM8K.

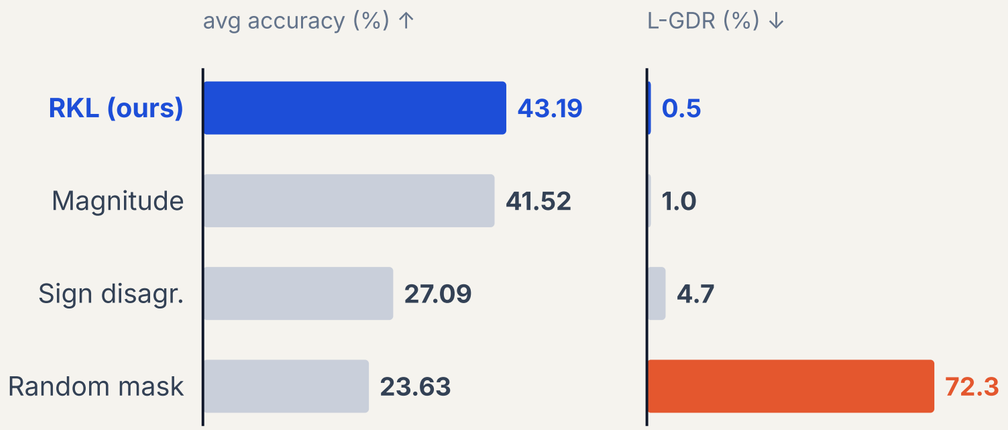
FOUR-TASK MEAN (%) VS. RECENT METHODS · PAPER TABLE 2



Highest mean on both pairs, though not on every task: a competitive capability–stability trade-off, not uniform dominance.

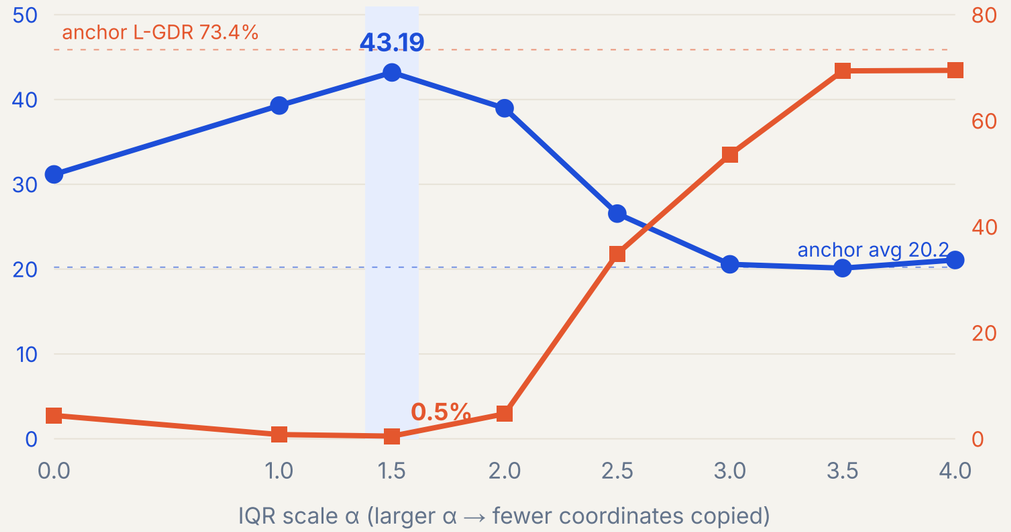
Both the criterion and the sparsity matter.

SELECTION CRITERION · SAME TENSOR-WISE IQR PROTOCOL, α = 1.5



Densities are close but not identical (RKL 5.43%, magnitude 5.98%). Random masks collapse on generation-heavy tasks: GSM8K 15.81 vs 57.26.

IQR SCALE α · AVG ACCURACY & L-GDR



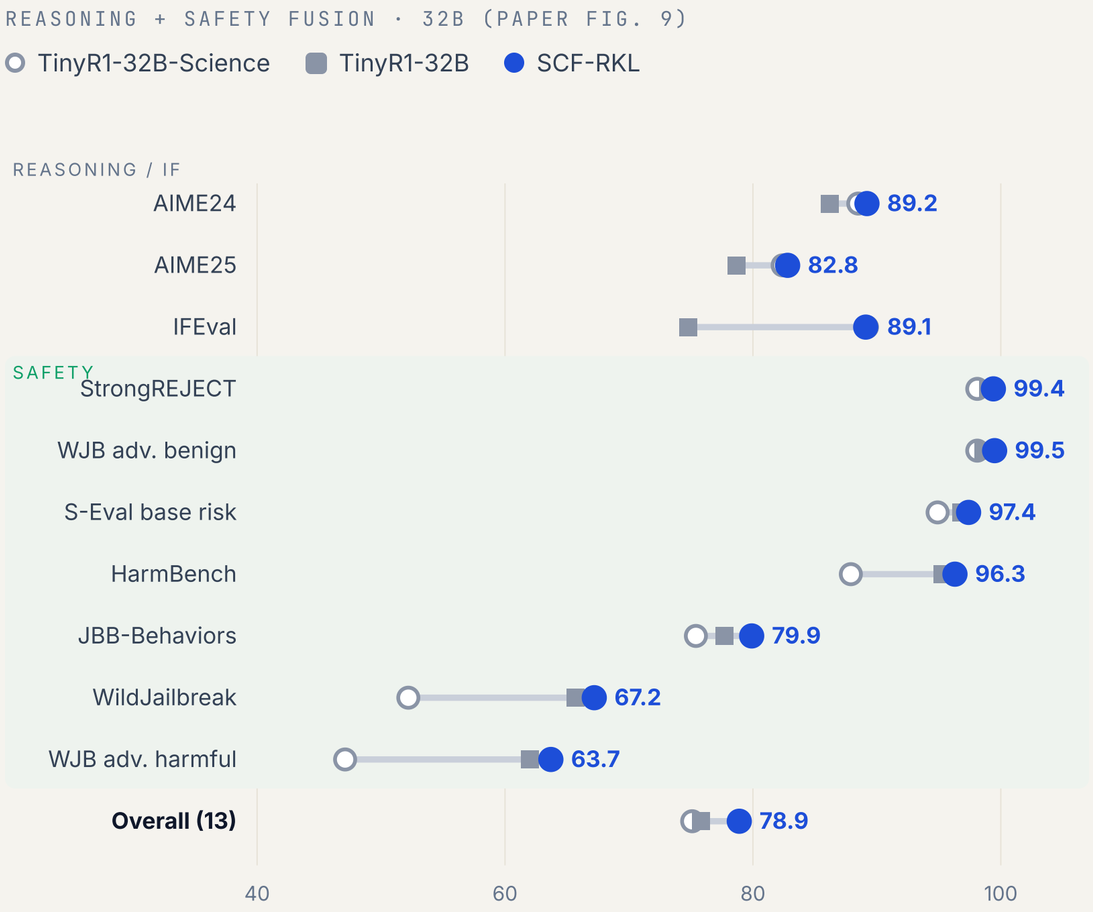
Too dense (α = 0, ~50% replaced) adds noise; too sparse (α ≥ 2.5) slides back to the anchor's 73% L-GDR. α = 1.5 is best on both.

Proxy is meaningful: row-level RKL correlates with intermediate-layer logit RKL for all 7 projection types (Spearman $\rho \geq 0.59$, $p < 10^{-3}$).

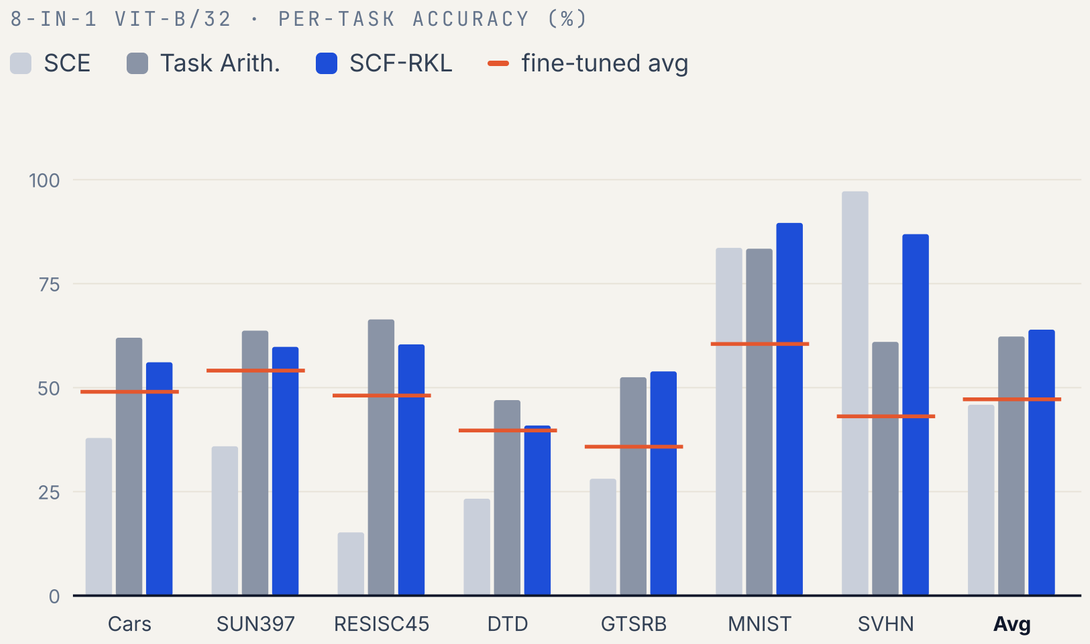
Anchor order matters: anchoring on MathCoder2 gives 43.7 vs 37.5 reversed; near-tied pairs can flip by task.

Safer and smarter. And it works for vision.

REASONING + SAFETY FUSION · 32B (PAPER FIG. 9)



8-IN-1 VIT-B/32 · PER-TASK ACCURACY (%)



78.9

overall vs 75.1 / 75.8 for the parents; ≈ +3 pts on 7 safety benchmarks

60.14

8-task macro avg incl. EuroSAT, vs 58.25 (TA) and 41.46 (SCE)

Select, don’t average.

Diagnose

Generation dysregulation is a real, measurable failure of model merging.

Select

Reverse-KL saliency + an IQR threshold copy a sparse tail; no phantom values.

Stay stable

L-GDR < 1% with competitive accuracy from 7B to 32B, safety and vision.

Paper on arXiv

arxiv.org/abs/2602.11717

Thank you!

Limitations: pairwise LLM merging needs a supplied anchor order · L-GDR depends on the stated decoding protocol · the row-softmax proxy is a heuristic.