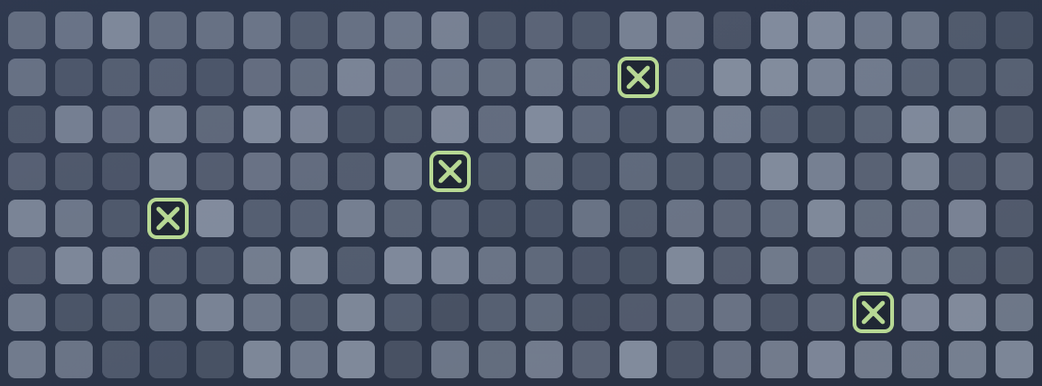
Fine-tune an LLM by deleting weights. Nothing else.

FROZEN WEIGHTS · LEARN ONLY WHICH TO SET TO 0



× = weight zeroed by the learned mask

WEIGHTS ZEROED

<0.05%

of all weights, and it matches full fine-tuning at the largest scale tested

THE PAPER

The Surprising Effectiveness of Deleting Weights in LLM Post-Training

Lu, Yang, Lasby, Pote, Ioannou, Ren · NYU, University of Calgary · NeurIPS 2026

THE IDEA

The adapter is a bitmask: keep or zero, one bit per weight.

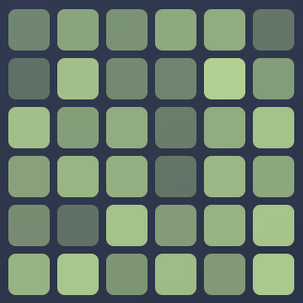


Mohammad Ashfaq Ur
AI Research Notes

Can a model learn new skills if the only edit allowed is deletion?

Post-training normally writes new numbers into the network. This paper asks how little has to change: not new values, just which pretrained weights to switch off.

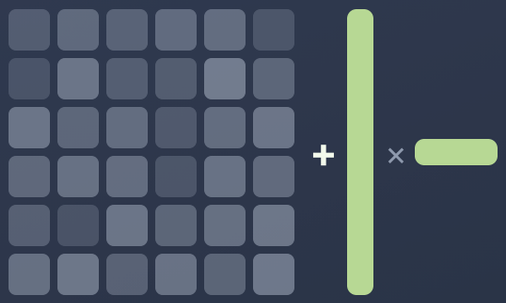
FULL FINE-TUNING



Any weight can take any new value.

`all params`

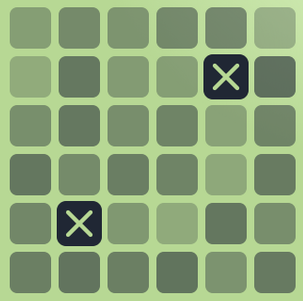
LORA



Frozen W plus a small trainable low-rank update B·A.

`new values, low rank`

DELETION TUNING



Frozen W. The only allowed edit is setting a weight to 0.

`1 bit per weight`

Why it's plausible: prior work cited in the paper finds RL fine-tuning naturally updates only 5 to 30% of parameters.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.10866 · NeurIPS 2026

Learn a keep-or-zero mask on the gate projection

1

Freeze

Every pretrained weight W stays fixed

2

Score

One trainable logit G per gate-projection weight

3

Soft mask

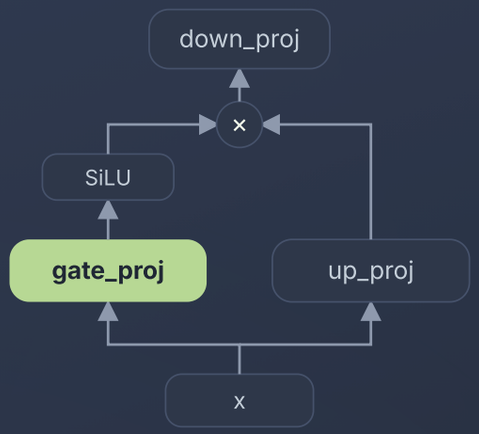
Train with $W \cdot \sigma(\alpha \cdot G)$; α ramps $10^3 \rightarrow 10^4$

4

Hard mask

Deploy $W \cdot 1[G > 0]$: keep or zero

WHERE IT LIVES · MLP BLOCK



DETAILS

- Gradients flow through the soft mask; the hard mask is used for evaluation and RL rollouts
- Starts from a Wanda pruning prior (~0.18% of gate weights zeroed); no fixed sparsity target
- Soft and hard masks end within **0.08 pts** of each other

THE ADAPTER

1 bit

per gate weight. Base weights are untouched.



Mohammad Ashfaq Ur
AI Research Notes

Deletion alone keeps up with full fine-tuning

EIGHT BACKBONES · 0.5B TO 30B · TWO FAMILIES

RL · GRPO	GSM8K: Llama 3.2 1B/3B, Llama 3 8B SimpleRL-Zoo maths: Qwen2.5 0.5B–14B, Qwen3-30B-A3B
SFT	Llama 3.2 1B/3B on 14 MMLU-Pro subjects and 23 BBEH subtasks
On-policy self-distillation	Qwen2.5-7B-Instruct on ToolAlpaca (tool use) and SciKnowEval Chemistry

Baselines: full fine-tuning (FFT), LoRA r = 1 and 4 on all seven projections, and random sparse fine-tuning.

PARITY WITH FFT

GRPO · SFT · SDFT

on par with FFT at the largest scale tested in each, up to a 30B mixture-of-experts model

BEYOND THE BASE MODEL

0 / 128

With GRPO, it solves some problems the base model failed in all 128 sampled attempts.

All claims from the abstract and Sections 1–5 of arXiv 2611.10866. Fewer than 0.05% of weights are deleted in every GRPO run.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev

arXiv 2611.10866 · NeurIPS 2026

A skill you can ship as a few megabytes, with less forgetting

ADAPTER SIZE · LOG SCALE



"est." = my arithmetic, not the paper's: Llama 3 8B has ~1.88B gate weights (1 bit each) and ~8.03B parameters (2 bytes each). 16 GB ÷ 8.4 MB ≈ 1,900×.

FORGETTING UNDER RL

−5.4

points on prior tasks: the worst drop for FFT, LoRA and random sparse fine-tuning. Deletion Tuning stays near the base model.

WHY IT MIGHT WORK

Its weight change is the closest to FFT's on stable rank, spectrum drift and principal-weight overlap (vs LoRA and random sparse). It also keeps forward KL on the new task low, which earlier work links to less forgetting.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev

arXiv 2611.10866 · NeurIPS 2026

Much of post-training may be switching off the right few weights.

CRITICAL PERSPECTIVE

- The headline is parity *at scale*: the claim is for the largest models tested in each setting
- Cheap to store, not to train: one logit per gate weight (~1.9B for an 8B model)
- Unstructured <0.05% sparsity gives no inference speed-up; only gate projections are masked
- Why deletion matches FFT is described (geometry, KL), not yet explained; no code linked yet

IF IT HOLDS UP

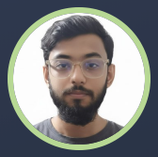
One base model could carry many skills as megabyte bitmasks, swapped in by deleting weights, with less forgetting than today's fine-tunes.

REFERENCES

arxiv.org/abs/2611.10866

arxiv.org/html/2611.10866

agenticlearning.ai/deleting-weights



Mohammad Ashfaq Ur

Follow for daily AI research breakdowns

ashfaqfardin.dev

