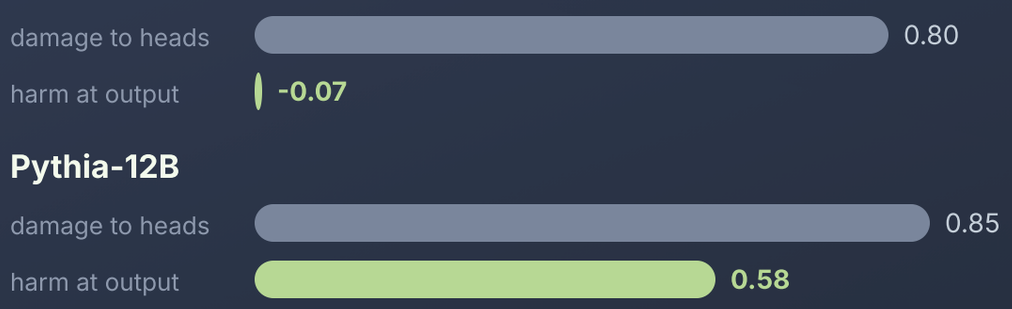
Compression that is harmless in a small model can be costly in a large one.

SAME DAMAGE TO RETRIEVAL HEADS · LOGITS



DAMAGE REACHING THE OUTPUT · PYTHIA-12B

68%

0.58 of 0.85 logits. At 410M the model repairs all of it (harm −0.07).

THE PAPER

When Bigger Becomes Brittle: Why Compression Hurts More as Language Models Scale

Gupta, Kukreja, Kumar, Mittal, Dabas, Sah, Cambria, Shah, Wang, Anand · IIIT Delhi, NTU, IIT Kanpur, SIT and others

THE IDEA

Separate the damage from the network's repair.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.07891 · preprint

A benchmark score mixes damage and repair.

When a compressed model scores almost the same, we can't tell if nothing broke, or something broke and the rest of the network covered for it.

DEFICIT · D

Direct contribution the damaged heads stop making to the answer.

REPAIR · D − H

What other components give back: suppression heads easing off, backup heads and MLPs stepping in.

HARM · H

How far the answer margin (correct name vs repeated name) actually falls.

EXAMPLE · PYTHIA-410M · 10 HEADS QUANTISED TO 4 BITS

1.70
logits lost by the heads

0.54
logits lost at the output

Why it matters

If repair shrinks with scale, a recipe that looks safe on a small model says little about a large one.



Mohammad Ashfaq Ur
AI Research Notes

An exact budget for where the damage goes

1

Pick a mapped task

Indirect-object identification: “When Mary and John went to the store, John gave a drink to …”

2

Find the heads

Score retrieval heads by where they attend; find copy-suppression heads by OV/QK tests

3

Damage them

Zero-ablation, simulated 4-bit quantisation of W_Q/K/V/O, or pruning masks

4

Do the accounting

Split the change in each component’s direct logit contribution

ACCOUNTING IDENTITY (EQ. 1)

$$H = D - R - B - N$$

- **D** deficit removed from targeted heads
- **R** release by copy-suppression heads
- **B** net backup response (other heads + MLPs)
- **N** final-normalisation term

MODELS

410M–12B

Six Pythia sizes, plus OPT 125M–6.7B and GPT-2 XL / Llama 3.1 8B case studies. 2,000 held-out prompts.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.07891 · preprint

From 2.8B up, most of the damage reaches the output



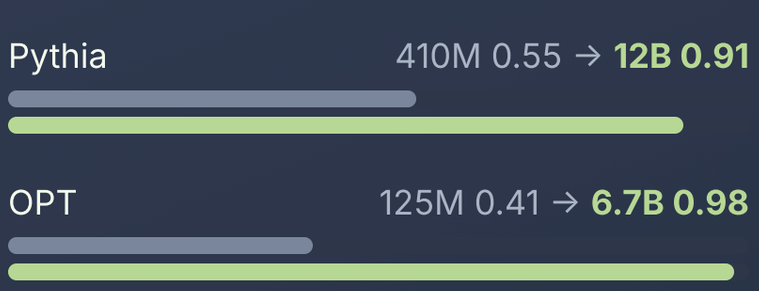
Negative H = output ends above its clean value. Attention response offsets ~80% of D − R at 410M vs ~9% at 12B. Source: Table 3a, arXiv 2611.07891.

Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.07891 · preprint

Same story under 4-bit quantisation, and small models carry more slack

SHARE OF DAMAGE REACHING OUTPUT · H / D



Simulated per-channel 4-bit quantisation of the target heads' weights. Intermediate sizes vary (Appendix D.1).

STANDING SUPPRESSION · SHARE OF CLEAN MARGIN

~25%
Pythia-410M

~5%
Pythia-12B

Copy-suppression heads hold the answer back a little. When retrieval weakens they ease off, and that reserve is what small models release.

Backup heads still react at scale: freezing 34 of them in Pythia-12B adds 0.30 logits of harm. Their net effect is small, and why remains untested.



Mohammad Ashfaq Ur
AI Research Notes

ashfaqfardin.dev
arXiv 2611.07891 · preprint

Compression tolerance doesn't transfer across scale.

CRITICAL PERSPECTIVE

- Preprint, not yet peer-reviewed; no code release listed
- One task: templated IOI prompts with single-token names
- Quantisation and pruning simulated in fp32; real deployment kernels not tested
- Checkpoint analysis is correlational; results depend on the readout convention

PRACTICAL IMPLICATION

Validate a quantisation or pruning recipe on the size you deploy. A small model shrugging it off is weak evidence.

REFERENCES

arxiv.org/abs/2611.07891

arxiv.org/html/2611.07891

github.com/TransformerLensOrg/TransformerLens



Mohammad Ashfaq Ur

Follow for daily AI research breakdowns

ashfaqfardin.dev

