Architecture, Not Scale: Circuit Localization in Large Language Models

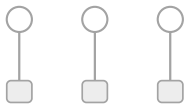
Sohan Venkatesh
Manipal Academy of Higher Education, Bengaluru

The Central Claim

Mechanistic interpretability assumes circuit analysis gets harder as models scale. Across three circuit types in Pythia and Qwen2.5, the **attention architecture predicts circuit geometry more reliably than parameter count.**

Is interpretability difficulty a fixed consequence of model size? No — it tracks architecture.

MHA · PYTHIA



Each query head owns its KV. **Diffuse** across tens–hundreds of heads.

GQA · QWEN2.5

Query heads share one KV. **One bottleneck** — ablate it, it breaks.

Why Architecture Drives Geometry

Four structural consequences of grouped-query attention

1. **KV Sharing Creates Bottlenecks**
   Ablating one shared KV head disrupts every query head assigned to it — a chokepoint with no analogue in MHA.

2. **Concentration, Not Diffusion**
   GQA circuits collapse into one or two heads across IOI, induction and factual recall. MHA spreads across tens to hundreds.

3. **Mechanistic Stability**
   The same GQA head dominates regardless of task difficulty or input distribution. MHA top heads shift between easy and hard regimes.

4. **Discrete Phase Transition**
   Above a critical scale, GQA factual-recall circuits collapse to a single layer-0 bottleneck rather than degrading gradually.

Experimental Design

Six models, 160M–7B · three circuit types · ablation via TransformerLens

PYTHIA · MHA: 160M · 1.4B · 6.9B

QWEN2.5 · GQA: 0.5B · 1.5B · 7B

IOI	Induction	Factual recall
Indirect object ID · logit diff (IO–S)	Repeated-token copy · ICL loss	Subject→object · accuracy

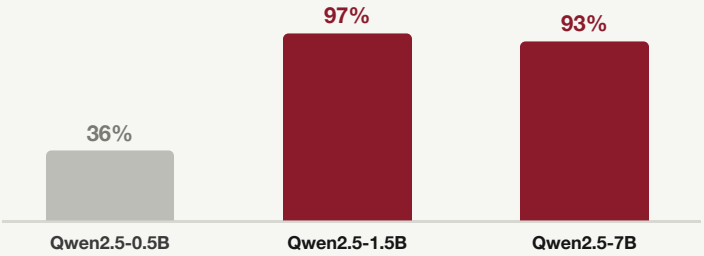
Circuit Concentration

Heads ablated to cause 80% damage — lower is more concentrated

	IOI	Induction	Factual
PYTHIA · MHA			
160M	5	>20	135
1.4B	2	22	50
6.9B	3	28	93
QWEN2.5 · GQA			
0.5B	1	2	1
1.5B	1	6	1
7B	1	6	1

A Discrete Phase Transition

Factual recall · accuracy lost when all layer-0 heads are ablated



Below 1.5B a backup pathway survives ablation; **above it** a single layer-0 KV head carries all factual recall.

Key Takeaways

- Attention **architecture predicts circuit geometry** more reliably than parameter count.
- GQA concentrates circuits into **one or two heads**; MHA spreads across tens to hundreds.
- Factual recall shows a **discrete phase transition** in GQA models above a critical scale.
- GQA circuits stay **stable across input conditions**; MHA circuits shift with difficulty.
- Most deployed frontier models already use GQA — **more interpretable than assumed.**

ACCEPTED · ICML 2026 MECHANISTIC INTERPRETABILITY WORKSHOP

Full paper & code
arXiv: 2605.08853