The Unembedding Bottleneck: A Mechanistic Analysis of Single Digit Counting in LLMs

Satwik Sunnam[1], Raghav Magazine[1], Vatsalya Singh[2],Lavanya Kotha[3], Xingjian Li[1], Min Xu[1]
Carnegie Mellon University[1], Indian Institute of Technology Guwahati[2], Independent Researcher[3]

Key Insight

The Unembedding Bottleneck

Counting failures in LLMs arise not from the absence of count information, but from a geometric misalignment between the count subspace and the unembedding matrix

The Problem

LLMs fail at elementary counting despite strong reasoning abilities.

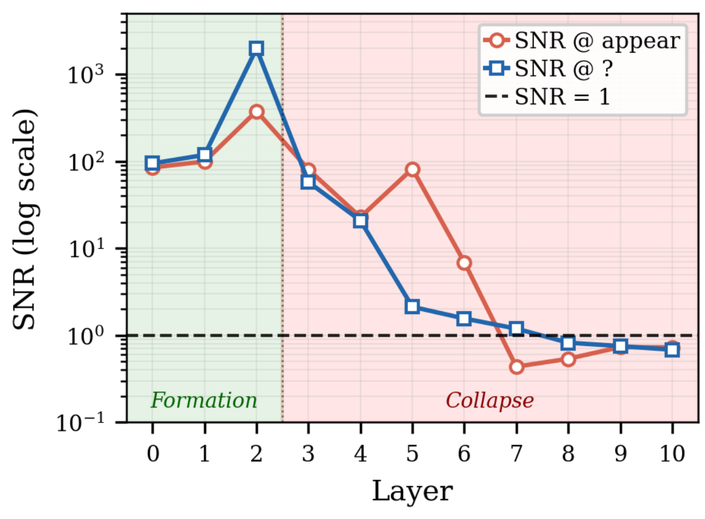
Why?

The failure is **geometric**, not representational.

Methodology

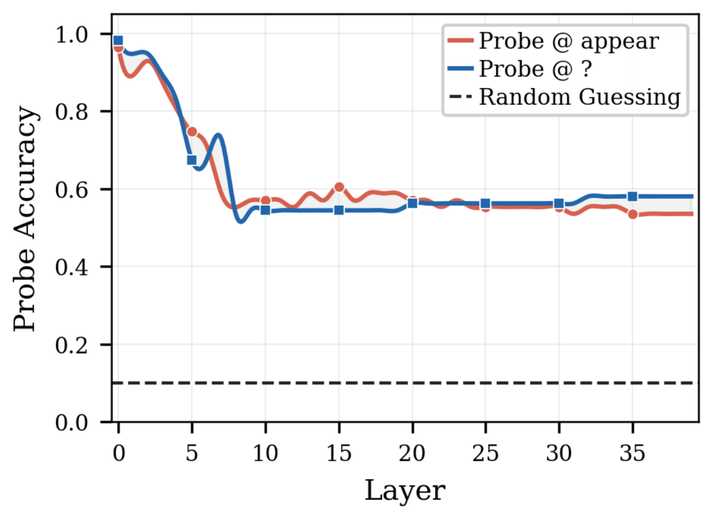
1. **SNR Analysis**: It gives a measure of the concentration of count signal across model layers
2. **Linear Probes**: It tells about the absence/presence of count signal and also serves as initialization for LoRA
3. **Probe-Informed LoRA**: Causal Validation of W_U bottleneck

SNR Across Layers



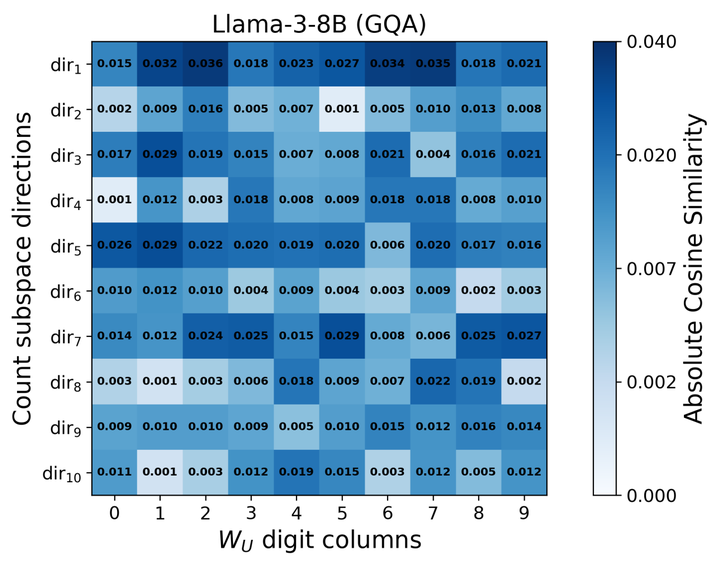
Count signal peaks early, collapses in deeper layers.

Probe Persistence



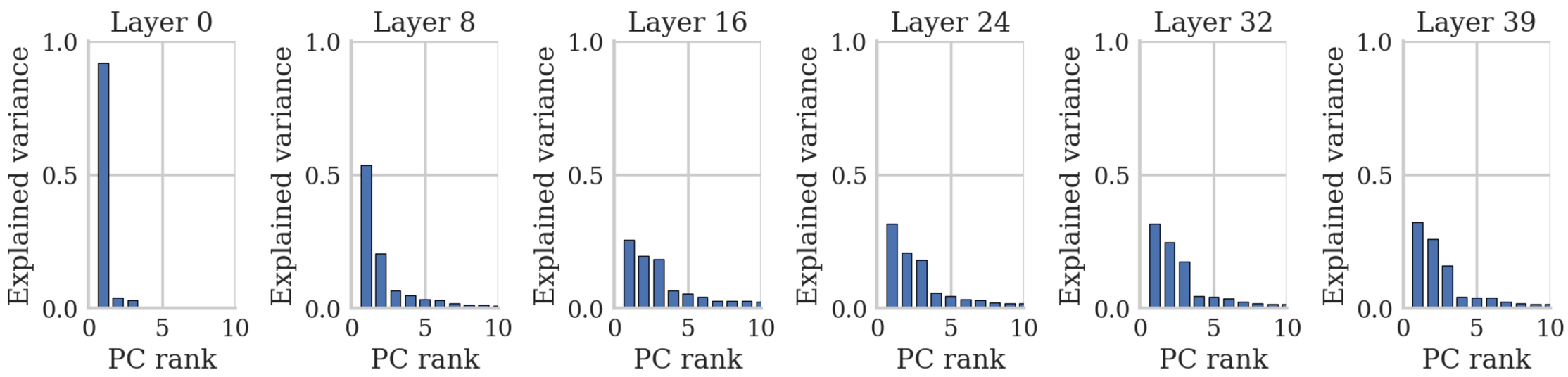
Linear probes show count signal is **dispersed**, not lost.

W_U Orthogonality



Probe directions $\perp$ digit-token columns of W_U.

Principal Component Dispersion

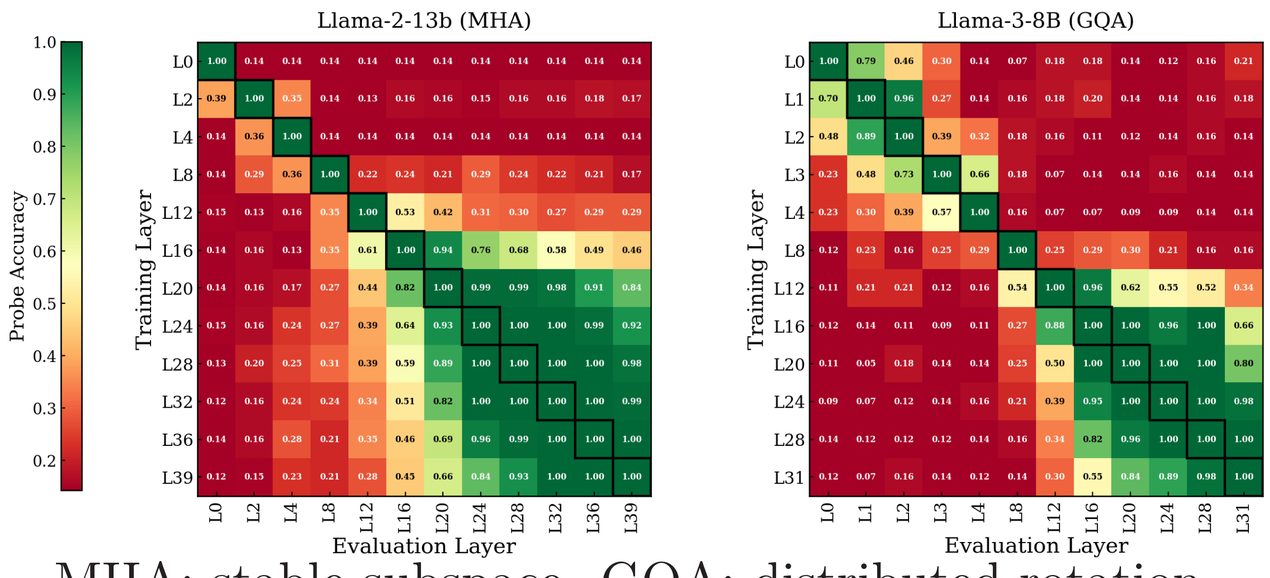


Count signal spreads across various principal component directions in deeper layers.

Causal Test

We initialize the W_U matrix with probe directions and only fine-tune it using our synthetic dataset with a rank-4 LoRA.
We update only $\sim$0.001% of model parameters during LoRA finetuning
We observe that **+80%** accuracy is recovered on single digit counting tasks

Cross-Layer Transfer



MHA: stable subspace. GQA: distributed rotation.