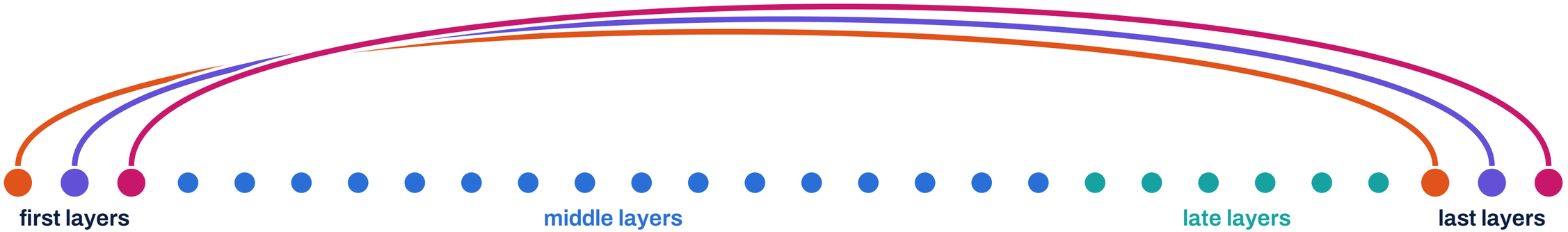
BRAINDECK

NeurIPS 2026
Poster Session 1, Tue 8 Dec, Hall 1-4, Sydney

Beyond Adjacent Layers

Graph-Guided Layer Fusion for Compressing Large Language Models

Muhammad Usama and **Yeoreum Jung**, Braindeck Inc, Seoul, Republic of Korea

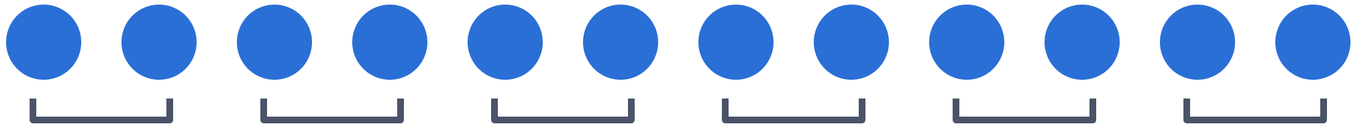


Which layers can we remove?

Drop the weakest layers



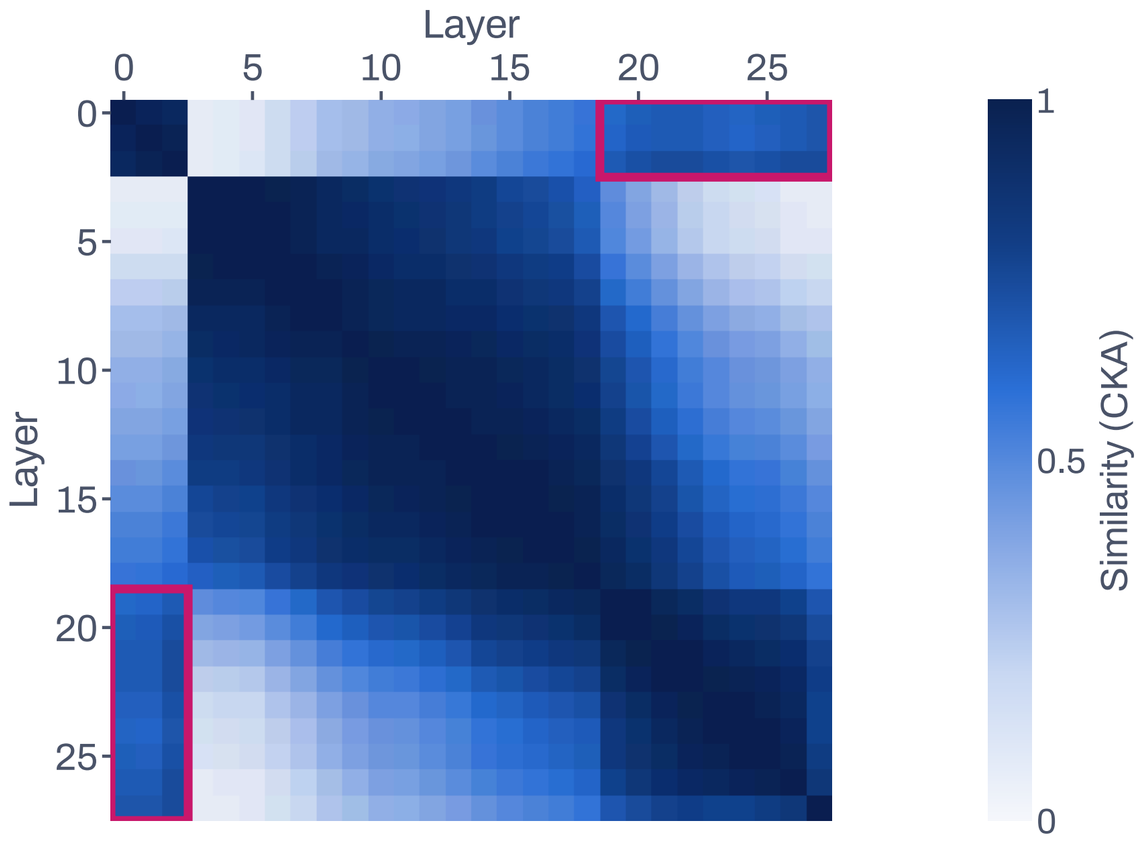
Merge neighbours



Both assume redundant layers sit next to each other.

Far-apart layers can be near-twins

Qwen2.5-7B, darker means more similar



Layer 2 vs. 3
neighbours

0.06

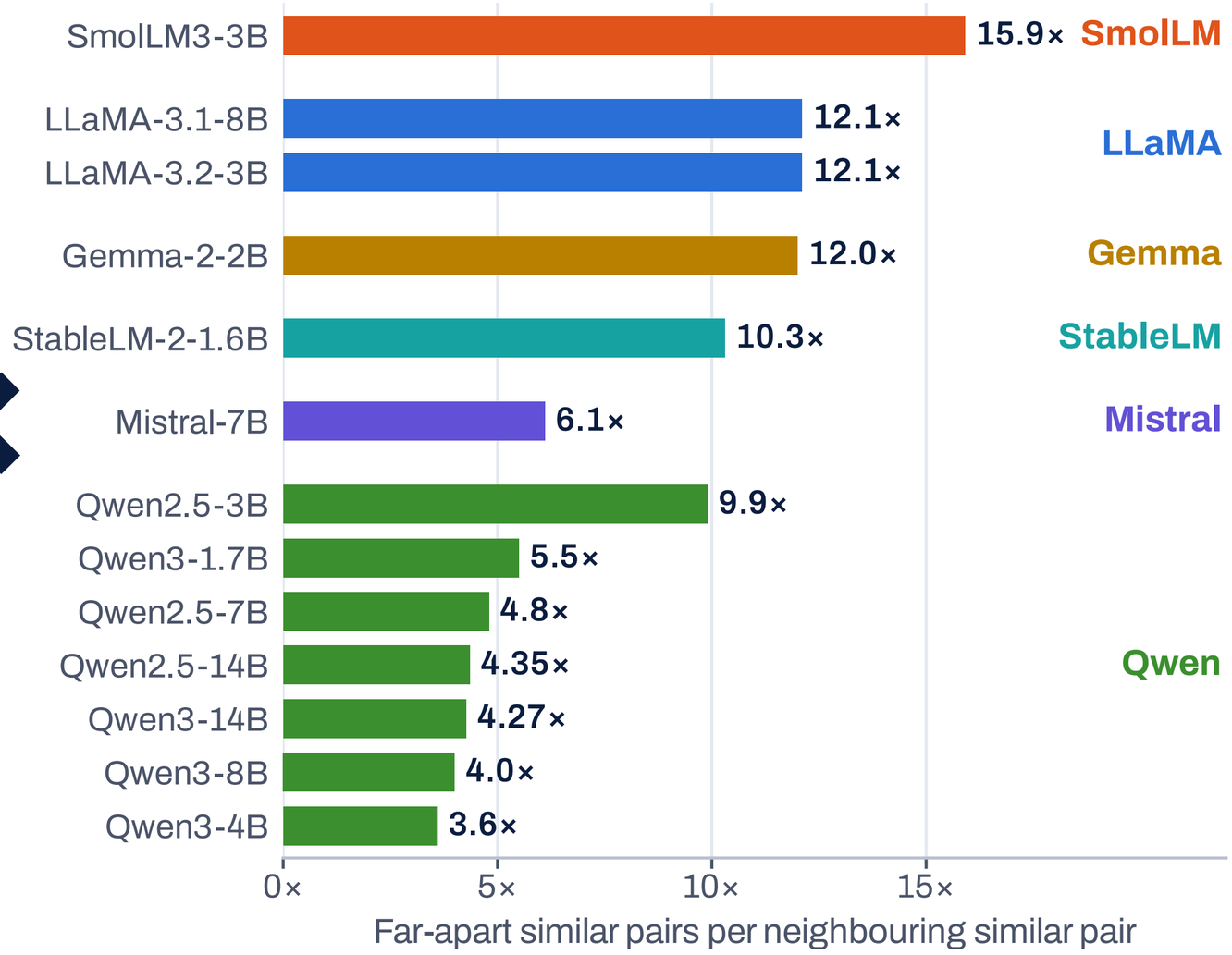
Layer 2 vs. 27
opposite ends

0.74

The pattern holds in all 13 models

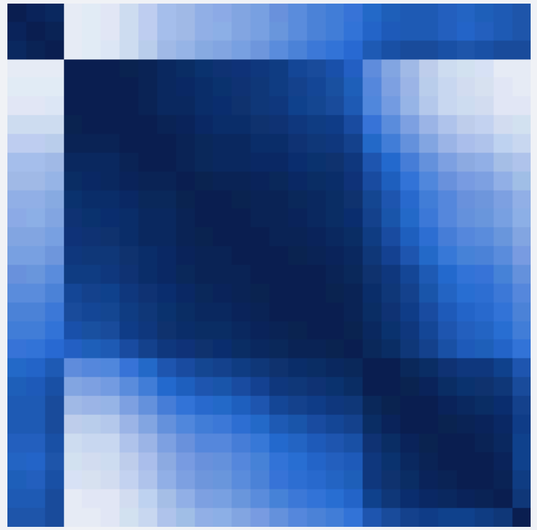
3.6–15.9×

more similar pairs **far apart** than side by side

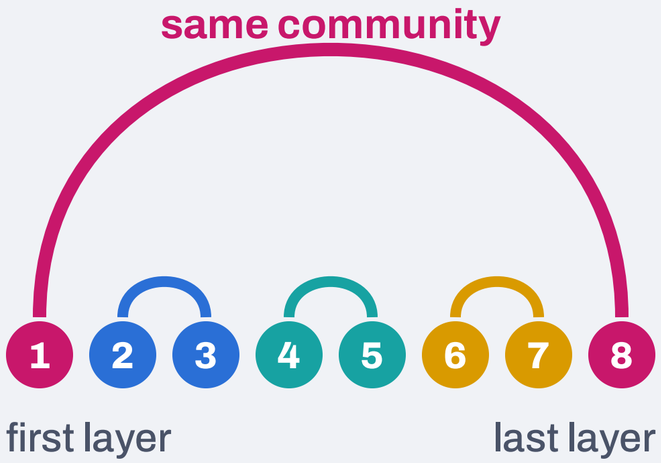


FCM in three steps

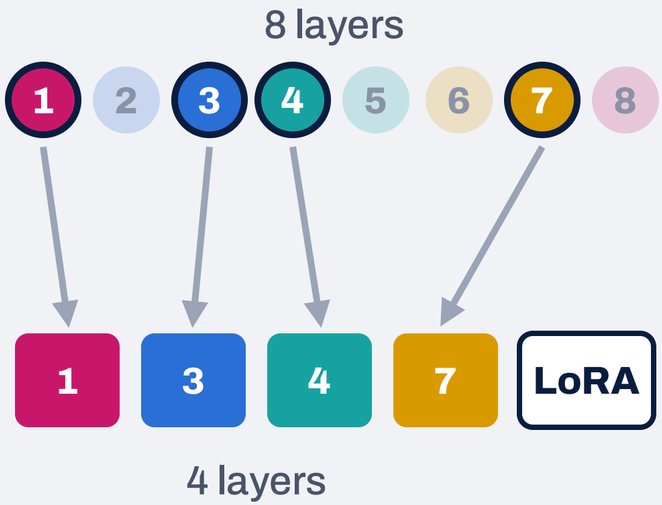
1 Measure similarity



2 Find communities



3 Keep the leaders



Which layer to keep matters

LLaMA-3.1-8B, perplexity (lower is better)

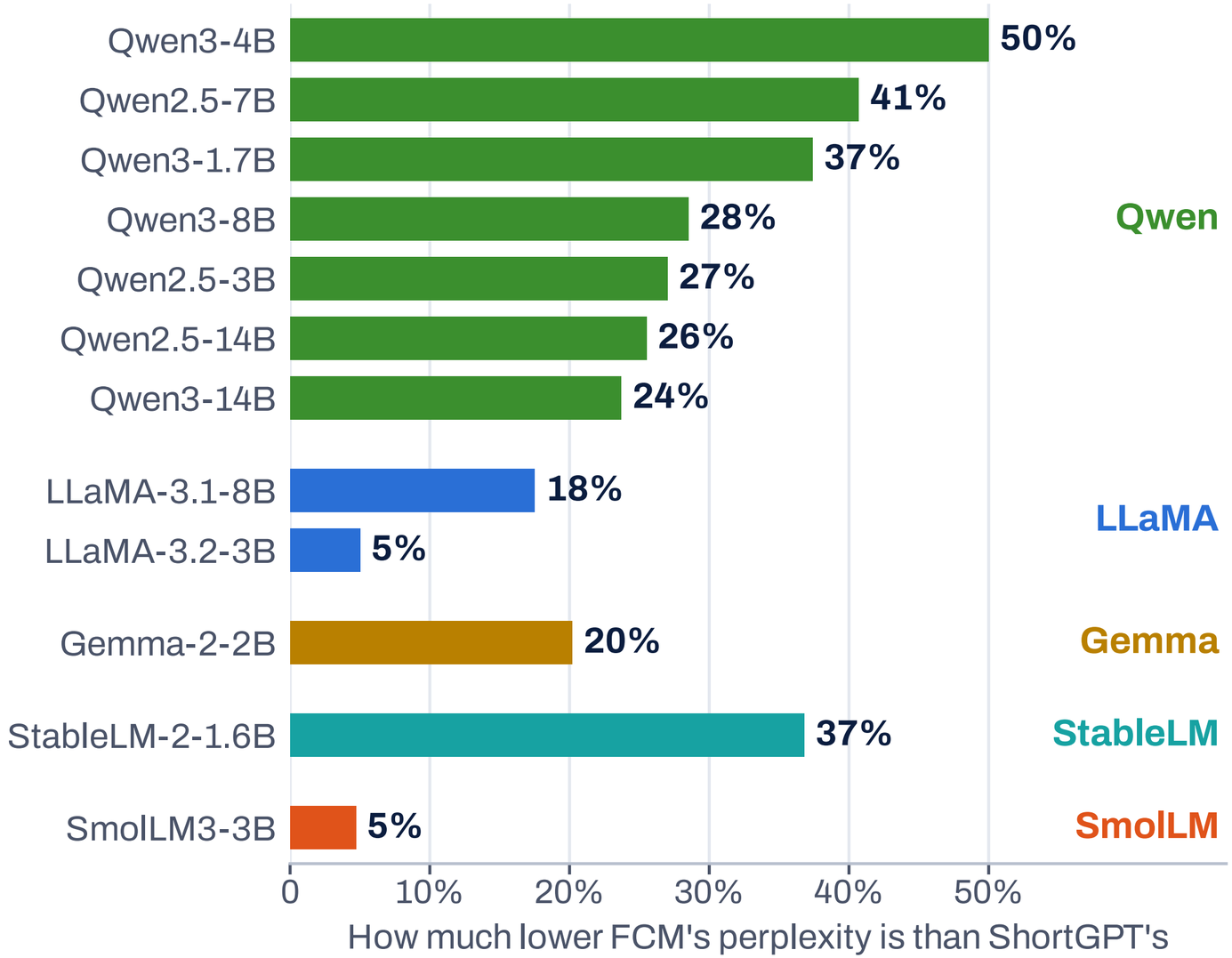
Most typical layer

450

Most active layer

35

FCM beats ShortGPT on 12 of 13 models



10 of 13

statistically significant

50%

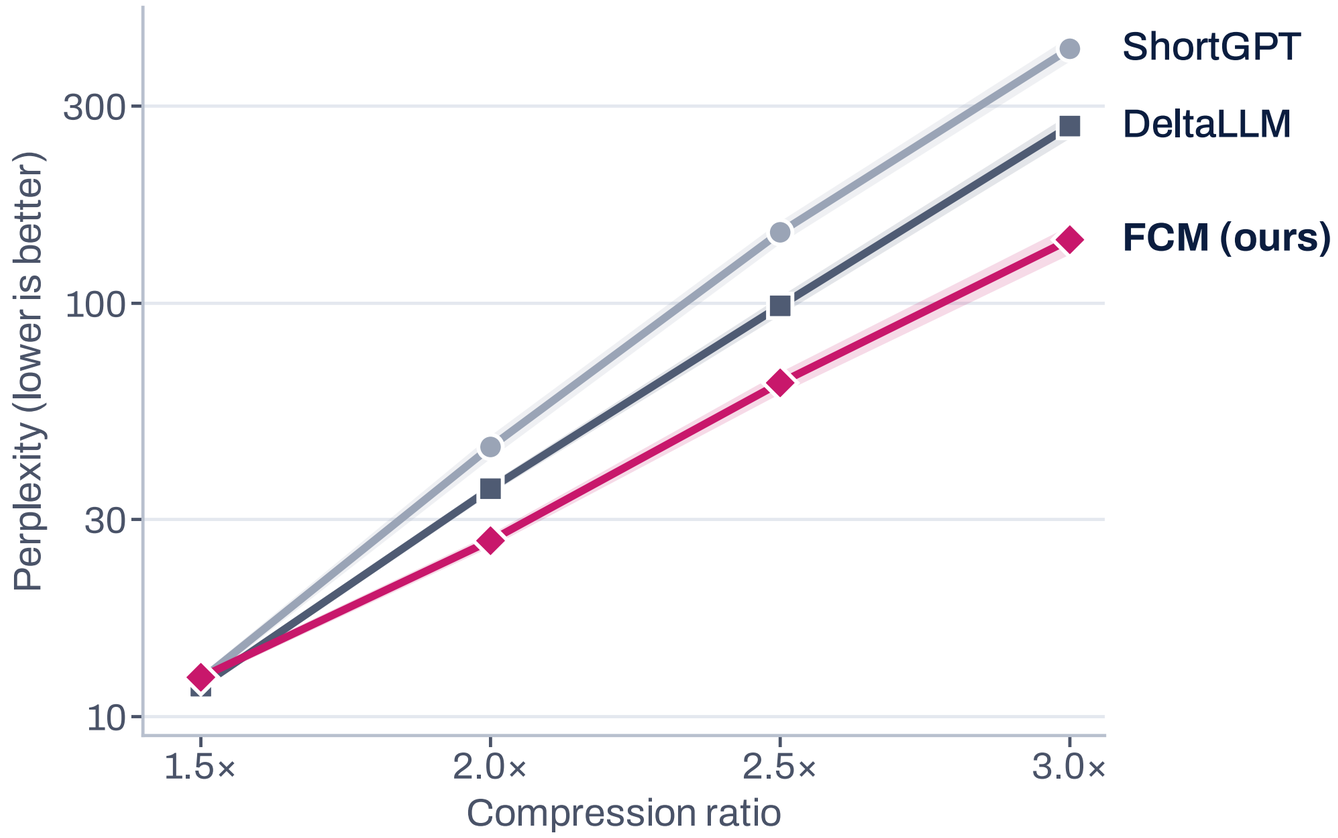
biggest gain, Qwen3-4B

10 of 13

best vs. FlattenGPT, SWM, DeltaLLM

The gap grows as we compress harder

Qwen2.5-7B, 2× removes half the layers, 3× two thirds



1/2

FCM's perplexity vs. the best baseline at 3×

Better on real tasks, and faster

Qwen2.5-7B, half the layers removed

	ShortGPT	FCM
Reasoning accuracy, %	49.3	55.5
Coding (HumanEval), %	18.5	29.8
Repetition loops	24%	2%

Smaller, faster

Parameters	–46%
Peak memory	–42%
Decoding speed	+84%

Takeaway

Redundancy in large language models is global, not local.

Look at the whole similarity graph and keep one strong layer per group.

Paper and code

usama1002.github.io

github.com/Usama1002/fcm-layer-fusion

Thank you

Visit our poster

Poster Session 1, Hall 1-4, Sydney

Tue 8 Dec 2026, 10:00 AM to 1:00 PM AEDT

BRAINDECK