Programs-of-Layers in LLMs through the Lens of Cortical Areas *(under review)*

SWAMBIT
SHARED WORKSPACE AND MEMORY FOR BRAIN-INSPIRED TRANSFORMERS

Justus Westerhoff,[1], Stephan Olbrich*,[1], Hatem Oraby[2], Walter Senn[3], Matthew Evan Larkum[2], Felix Gers[1]*
[1]Berliner Hochschule für Technik (BHT), [2]Humboldt-Universität zu Berlin, [3]University of Bern
**shared first authorship, random order*

NI:KI INITIATIVE

Project page, paper, code, ... and more!

Can we improve transformers by skipping and duplicating their layers?

1. Pre-Study — Can transformer layers cooperate out of their usual order?

Question: Point A has coordinates (x,6). When Point A is reflected over the y-axis it lands on Point B. What is the sum of the four coordinate values of points A and B?

(sample from DART-Math, difficulty tier 1)

1 → 2 → 3 → 4 → 5 → 6 → 7

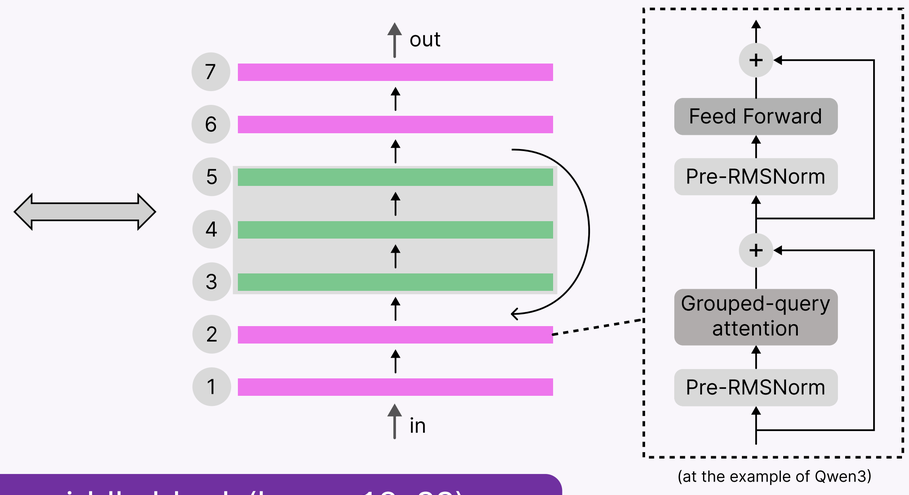
The LLM as trained in its usual order ("identity").

✗ *Answer: 6*

LLM layers executed in a modified order.

1 → 2 → 3 → 4 → 5 → 3 → 4 → 5 → 6 → 7

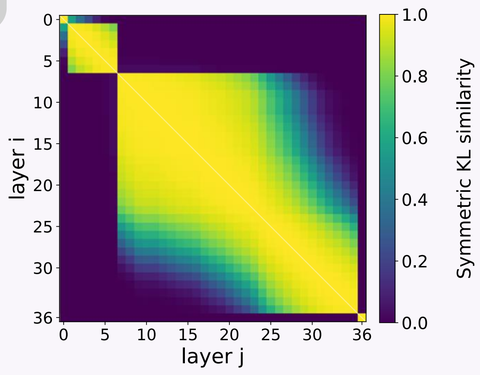
✓ *Answer: 12*



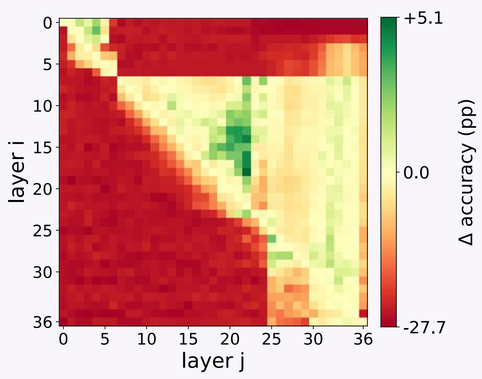
Inspiration: David Noel Ng, *LLM Neuroanatomy: How I Topped the LLM Leaderboard Without Changing a Single Weight, March 2026*, https://dnhkng.github.io/posts/rys/

Qwen3-8B improves on a 6-domain subset of MMLU-Pro when repeating a middle block (layers 18–22).
The **representational** and **behavioral** patterns in all three figures agree on two boundaries: layer 6 and layer 25.

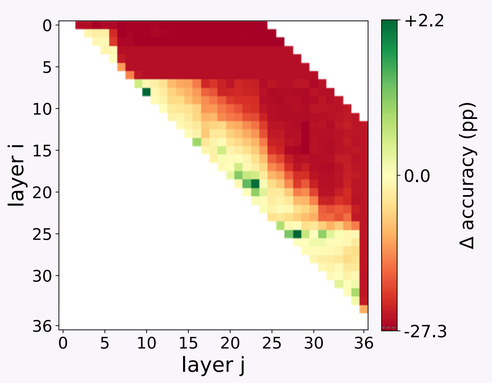
Symmetric KL similarity between layer i's and layer j's output token distributions.



Re-running block $[i, j)$ a second time (repeat, upper triangle) or removing it (skip, lower triangle).



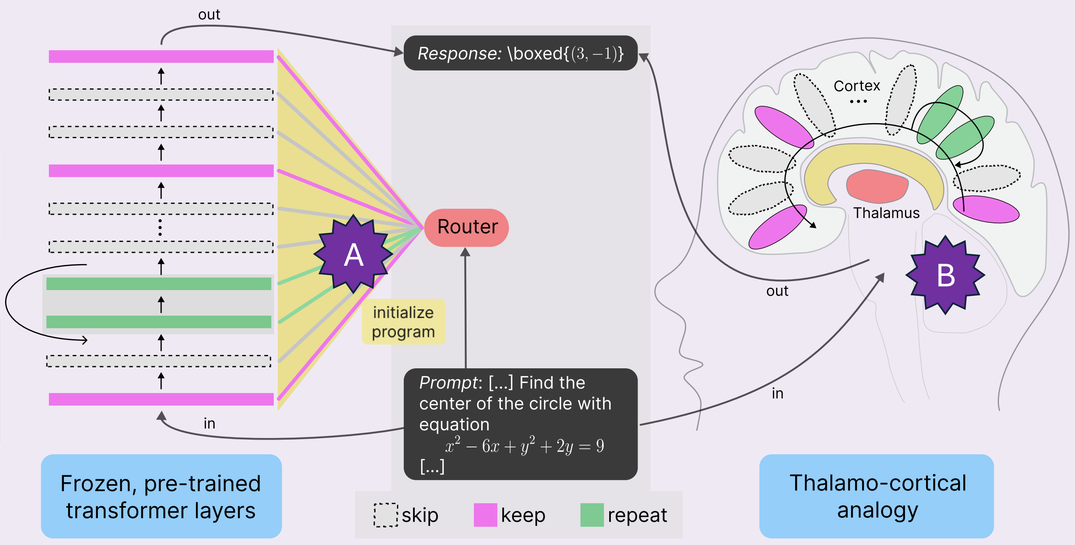
Running block $[i, j)$ for $j - i$ times, in parallel.



2. Cortical Analogy — Cortical areas as transformer layers.

A Li et al. [1] introduce program-of-layers (**PoLar**), using layers of transformers as a **library of functions** rather than a fixed sequence. They further train a lightweight **router**, that predicts programs for unknown sequences.

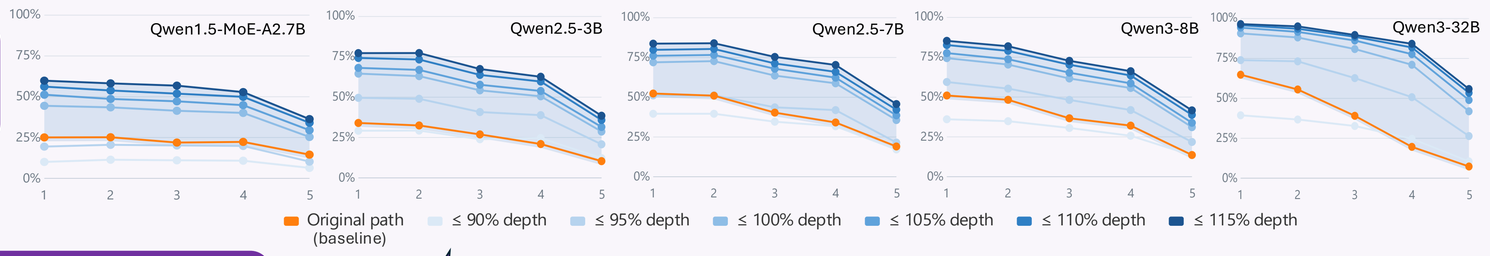
B **Functionally**, we see an analogy to the thalamus routing computation across cortical areas. That cortical areas realize attention-like computation is described in Babu et al. [2]. Larkum [3] argues that a column's architecture and its interaction with the thalamus let the brain integrate context into local computation and distribute the context-sensitive output around the cortex.



[1] Ziyue Li, Yang Li, and Tianyi Zhou. Skip a layer or loop it? learning program-of-layers in LLMs. In Forty-third International Conference on Machine Learning, 2026.
[2] Vasanth Kumar Babu, Amo Granier, Timothée Proix, Marmaduke Woodman, Viktor Jirsa, Titas Balvocius, Ausra Saudargiene, and Walter Senn. Cortico-subcortical multi-head self-attention as a substrate for cognitive performance. bioRxiv, 2026.
[3] Matthew Larkum. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. Trends in Neurosciences, 36(3):141–151, March 2013.

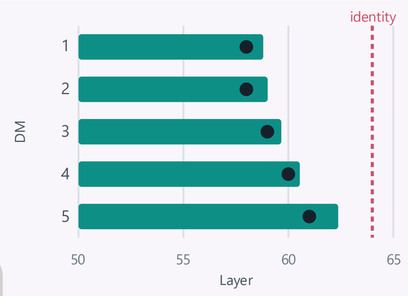
3. Programs-of-Layers — Searching programs: Running a Monte-Carlo Tree Search for every sample (here, focused on DART-Math, split into 5 difficulties).

I Baseline is beatable...

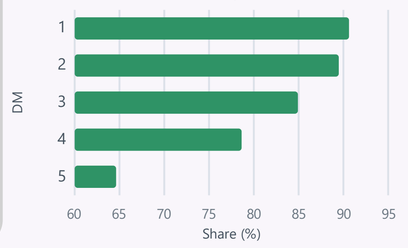


II ... even with shorter programs (Occam's Razor)!

Mean and median of the shortest program's executed length (Qwen3-32B).

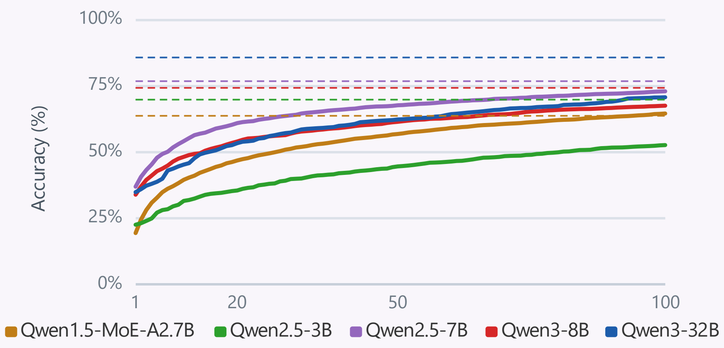


Share of programs admitting a shorter valid program (Qwen3-32B).

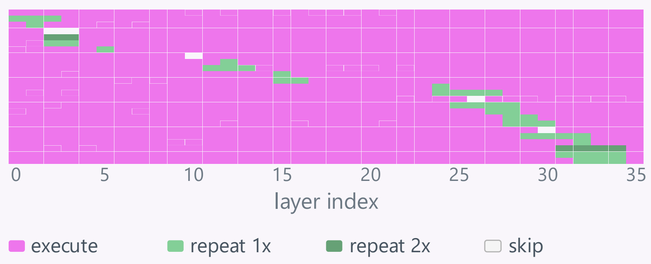


III A few programs solve a lot!

Coverage of the top-k programs (dashed = MCTS baseline).

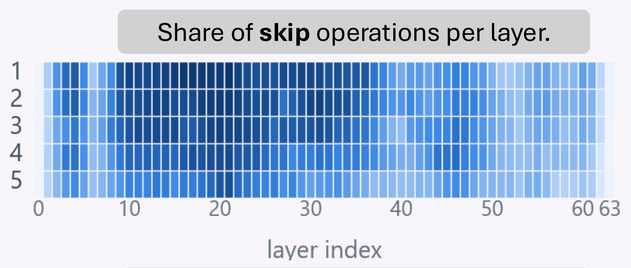


Top-25 programs ordered by edit position (Qwen3-8B).

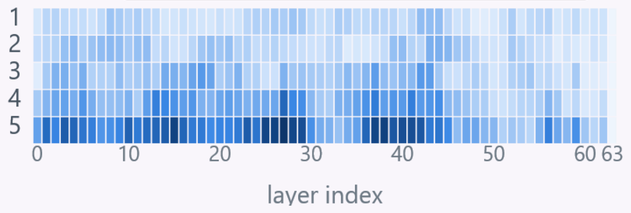


IV More repetitions for harder programs!

Share of **skip** operations per layer.



Share of **repeat** operations per layer.



Shortest valid programs across difficulties for Qwen3-32B.

Take-home message
There is a functional similarity between transformer layers and cortical columns.

There's more! In our paper, we give more detail about the **MCTS**, go deeper into **program robustness**, training a **router** on top of found programs, looping multiple times, and check the method on another dataset.

Very related Major related works are LayerDrop, LaCo, Mixture-of-Depths, ShortGPT, LayerSkip, FlexiDepth, MindSkip, GateSkip, **DR.LLM**, and, of course, **PoLar** [1] itself.

With funding from the: Federal Ministry of Research, Technology and Space
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
Our work is funded by the Federal Ministry of Research, Technology and Space under grant number 01GQ2505A and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – FIP-12 – Project-ID 528483508. Responsibility for the content of this publication lies with the author.

INITIATIVE FOR APPLIED FOUNDATION MODEL RESEARCH
NI:KI INITIATIVE
SWAMBIT SHARED WORKSPACE AND MEMORY FOR BRAIN-INSPIRED TRANSFORMERS
BHT Berliner Hochschule für Technik
HUMBOLDT-UNIVERSITÄT ZU BERLIN
UNIVERSITÄT BERN