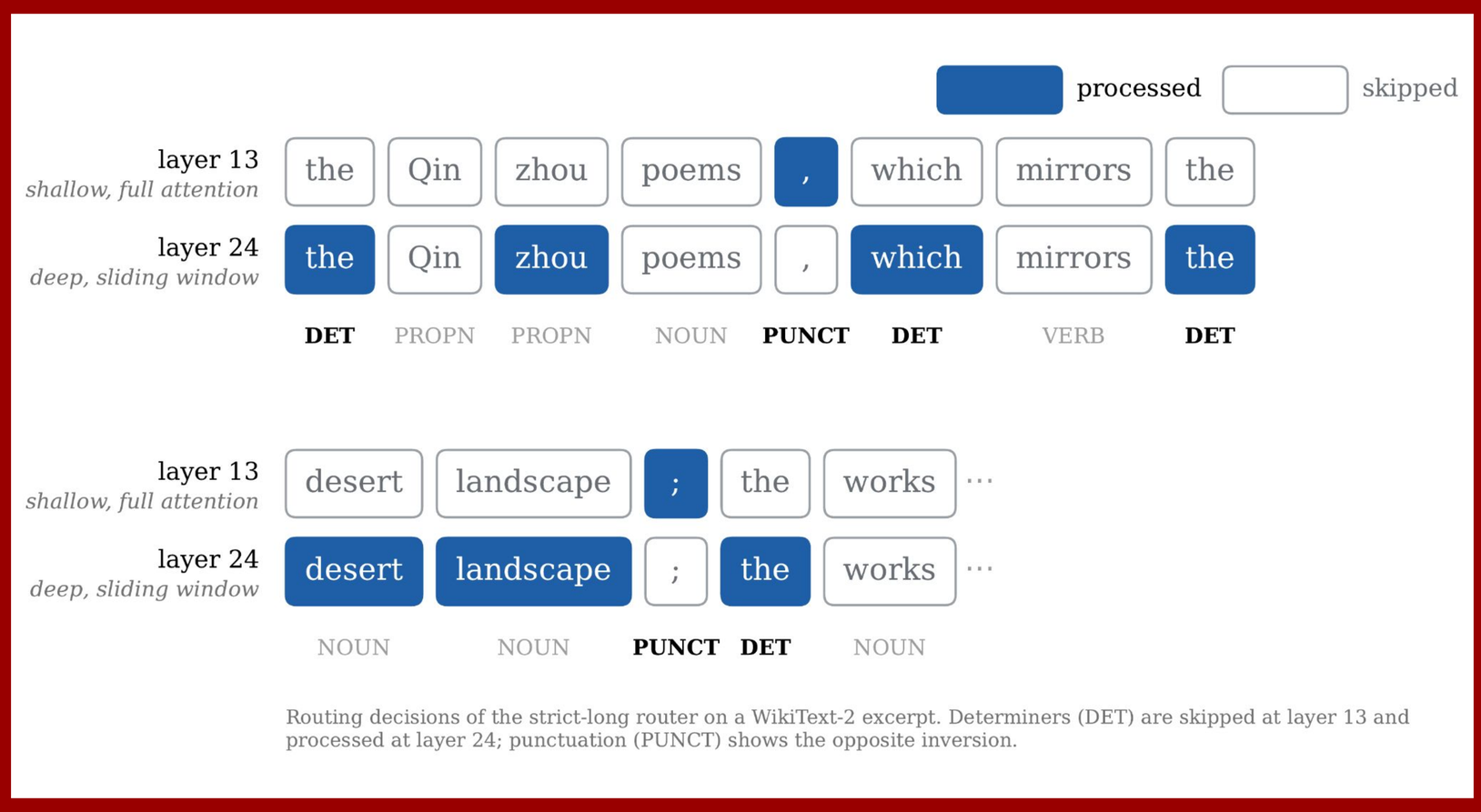
MoD routers learn layer-dependent, syntax-sensitive routing rules



Routing decisions of the strict-long router on a WikiText-2 excerpt. Determiners (DET) are skipped at layer 13 and processed at layer 24; punctuation (PUNCT) shows the opposite inversion.

What Do Mixture-of-Depth Routers Learn? Routing Patterns in Gemma 2

Stein Pleiter, Ege Erdogan, Ana Lucic

University of Amsterdam

1 TL;DR

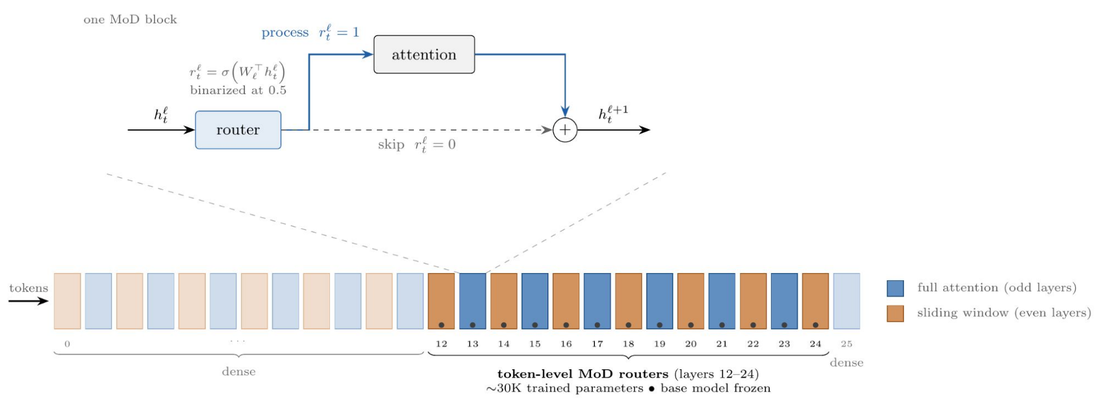
- We train token-level MoD routers on frozen Gemma 2 2B: to our knowledge, the first study of learned routing in an alternating-attention model.
- Routers consistently process more tokens at full-attention layers than sliding-window layers in shallow-middle layers; simple architectural explanations all fail.
- Routing encodes part of speech, with layer-conditional inversions.

2 Background / Motivating Question

- MoD adds a tiny per-layer router deciding, per token: process or skip the attention block.
- Prior work treats routers purely as efficiency mechanisms; their learned decisions have not been analyzed
- Modern models like Gemma 2 alternate full and sliding-window attention; unexplored for MoD.
- RQ1–3: Do routers route by attention type? Can simpler properties explain it? Do decisions encode syntax?

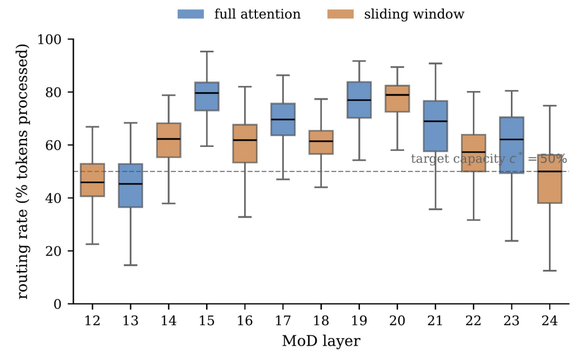
3 Setup: Router-Tuning on Gemma 2 2B

- Base model frozen; only routers trained (~30K params, layers 12–24). Any structure found must already exist in the pre-trained residual stream.
- Four checkpoints span capacity penalty $\lambda \in \{0.01, 0.1\} \times$ training scale (5K×1 vs 52K×3 Alpaca); evaluated on held-out WikiText-2 (~33K tokens).

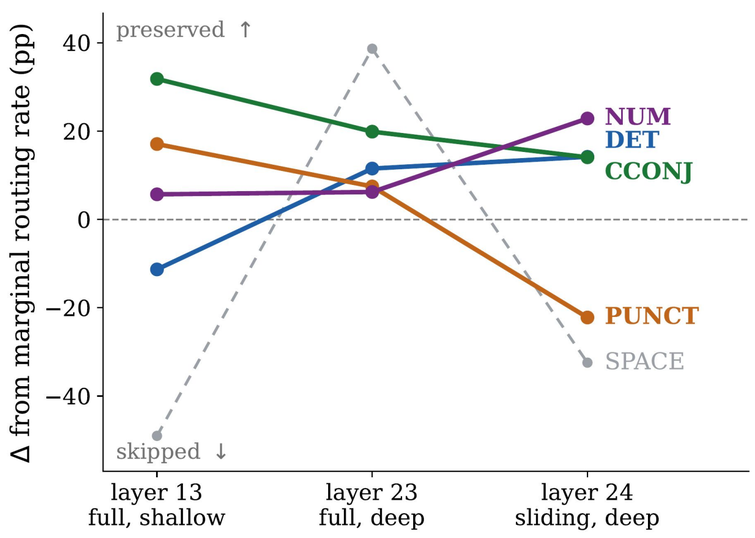


4 Routers prefer full attention, and simple explanations fail

- RQ1: All four checkpoints route more at full-attention layers in shallow-middle MoD; deep-MoD direction is training-dependent.
- RQ2: Not explained by layer norms (wrong direction, d = −3.9), per-token norm (≤11% of score variance), or ablation importance (p = 0.44; MLP control flat).



5 Routing encodes part of speech, with layer-conditional inversions



- Part of speech predicts routing at every target layer. Probe exceeds majority baseline by +6.6pp (L13) and +9.1pp (L24).
- The same category can be skipped at one layer and preserved at another; SPACE goes from 0% to 100% to 17% processed.

6 Limitations and Discussion

- Probes show correlation, not causation, residual streams encode POS densely (probe acc > 0.93).
- Parity confound: full attention ⇔ odd layers in Gemma 2; within-layer POS effects escape it.
- Next: a second alternating architecture, activation patching of POS directions, Gemma Scope per-feature analysis.

PAPER

https://openreview.net/forum?id=5HXw5MISXk

stein.pleiter@student.uva.nl