# Adaptive Layer Skipping during LLM Inference

Mgr. Gabriela Chutňáková  supervisor: Mgr. Vladimír Boža, PhD.

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS
Comenius University
Bratislava

## The idea

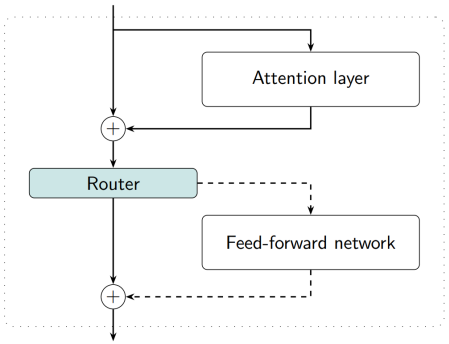
Large language models process every token through every layer, even when some computations have little effect on the result.

**What if the model could predict, token by token, which layers have little impact on the result?**

Instead of permanently removing layers, we let each token follow its own computational path.

## Gated LLM

A lightweight router gate is placed before the feed-forward network (FFN) in each LLM block.




For every token, the router makes a decision whether to **execute** or **skip** the FFN.

## Experimental setup

**Model:** SmolLM2-1.7B

**Calibration dataset:** C4

**Evaluation dataset:** WikiText-2

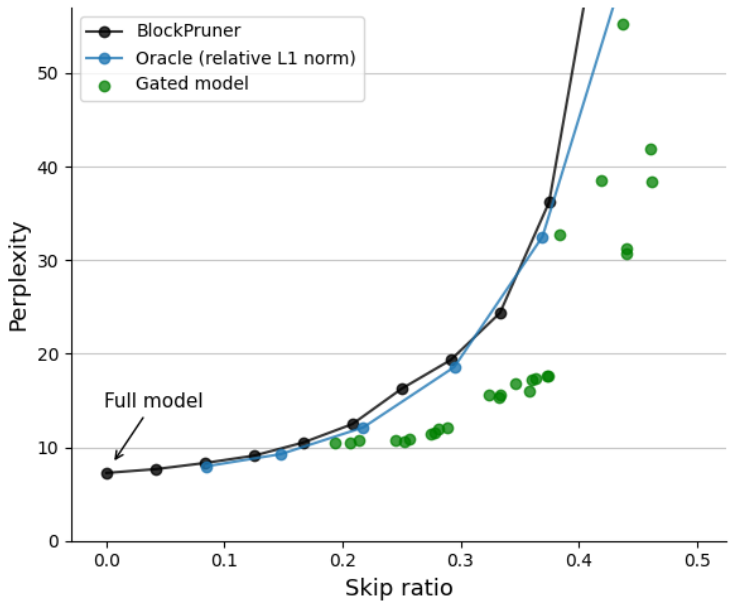
## How do routers learn?

The goal is to **reduce computation** while having the modified LLM **imitate** the full model.

The routers are trained jointly using **KL divergence** between the outputs of the modified and the full model. The original LLM parameters remain **unchanged** during training.

## Results

The figure compares gated models at different skip ratios, alongside baseline methods **BlockPruner** (static pruning) and **the oracle** (ideal magnitude-based routing).




On the WikiText-2 benchmark, the gated model **saves 33 % of FFN computations**, with perplexity increase only to **15.3** (compared to the full model's 7.3).

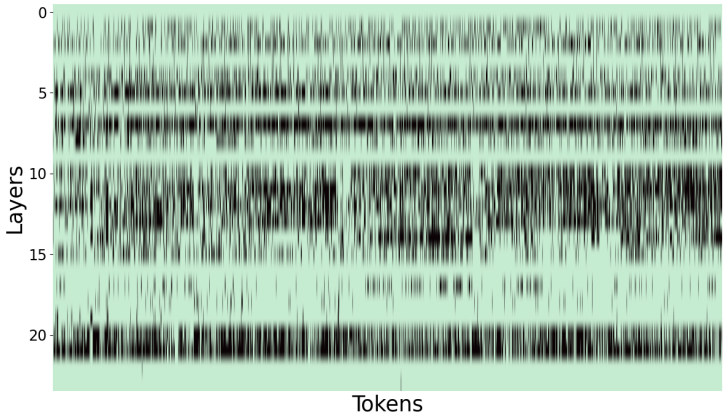
## Conclusions

Adaptive routing can remove a substantial amount of FFN computation while preserving much of the model quality.

Magnitude alone is not enough. A layer can have a small output while still being important to the model's final behaviour.

**Which layers are skipped?**

Black indicates that an FFN is skipped. Tokens move through the LLM from top to bottom.




Some layers are used almost all the time, while others can be skipped for more than half of the tokens.

## Future directions

The LLM used in our experiments is relatively small. **Larger LLMs** are likely to exhibit even more internal redundancy, which would provide more opportunities for layer skipping.

An attractive extension towards higher speedups is **an efficient GPU execution of layer skipping**. In practice, token-level skipping leads to irregular computation, which can be challenging on current GPUs.