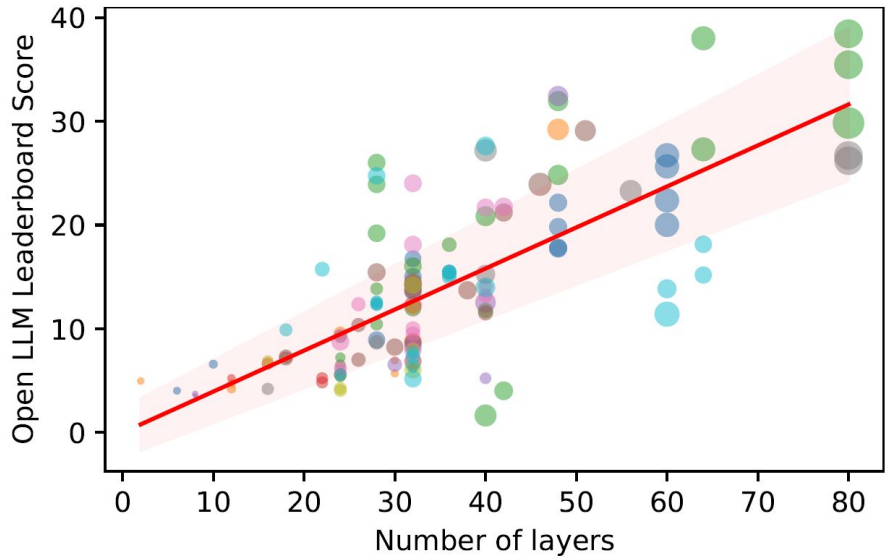
Do Language Models Use Their Depth Efficiently?

Daven Lim

Introduction

Why are LLMs so deep?

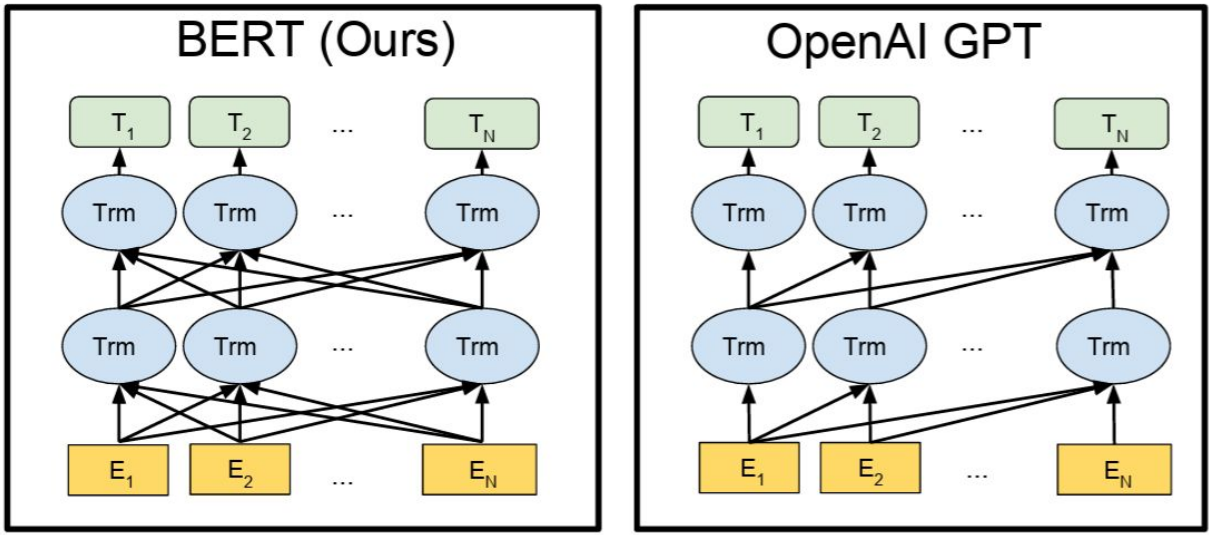


Colors: model families
Dot size: parameter count

Csordás, Róbert, Christopher D. Manning, and Christopher Potts. "Do language models use their depth efficiently?." arXiv preprint arXiv:2505.13898 (2025).

Introduction

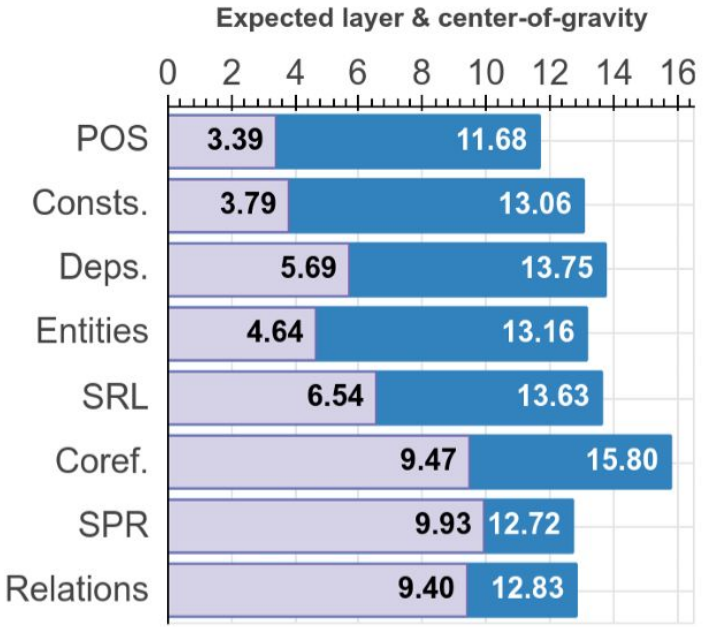
LLM Architectures



Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171-4186. 2019.

Introduction

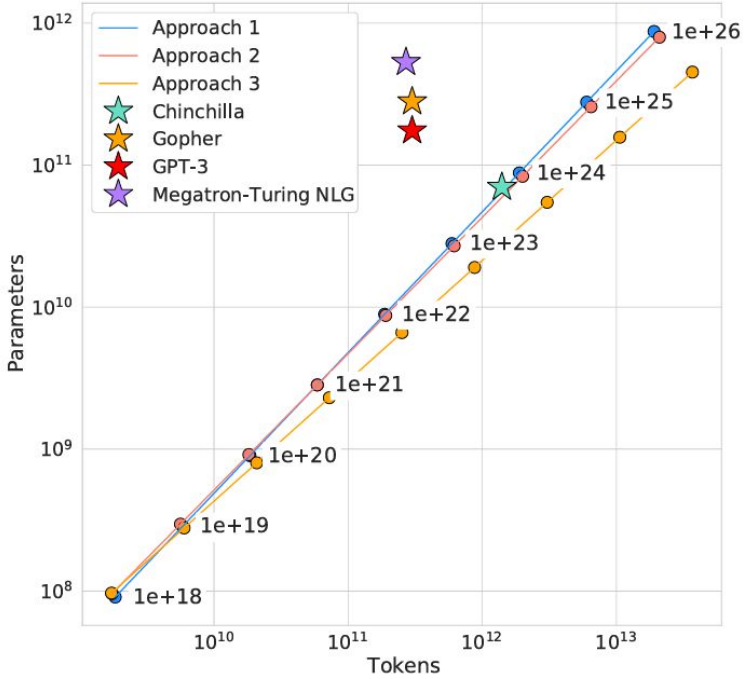
Composition in LLMs



Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline." In Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 4593-4601. 2019.

Introduction

Model Capacity Inefficiencies



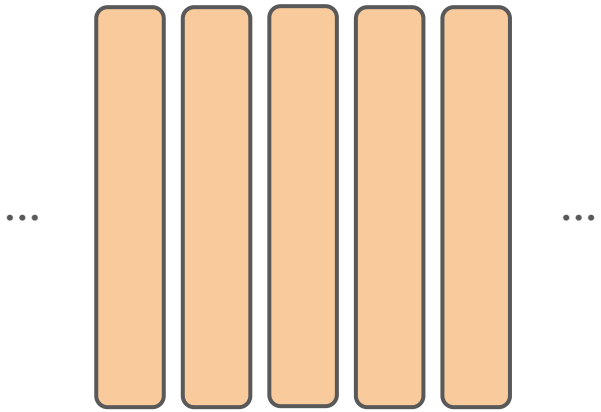
Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, D. D. L. Casas et al. "Training compute-optimal large language models." arXiv preprint arXiv:2203.15556 10 (2022).

Do Language Models Use Their Depth Efficiently?

Róbert Csordás **Christopher D. Manning** **Christopher Potts**
Stanford University, Stanford, CA, USA
rcsordas@cs.stanford.edu {manning,cgpotts}@stanford.edu

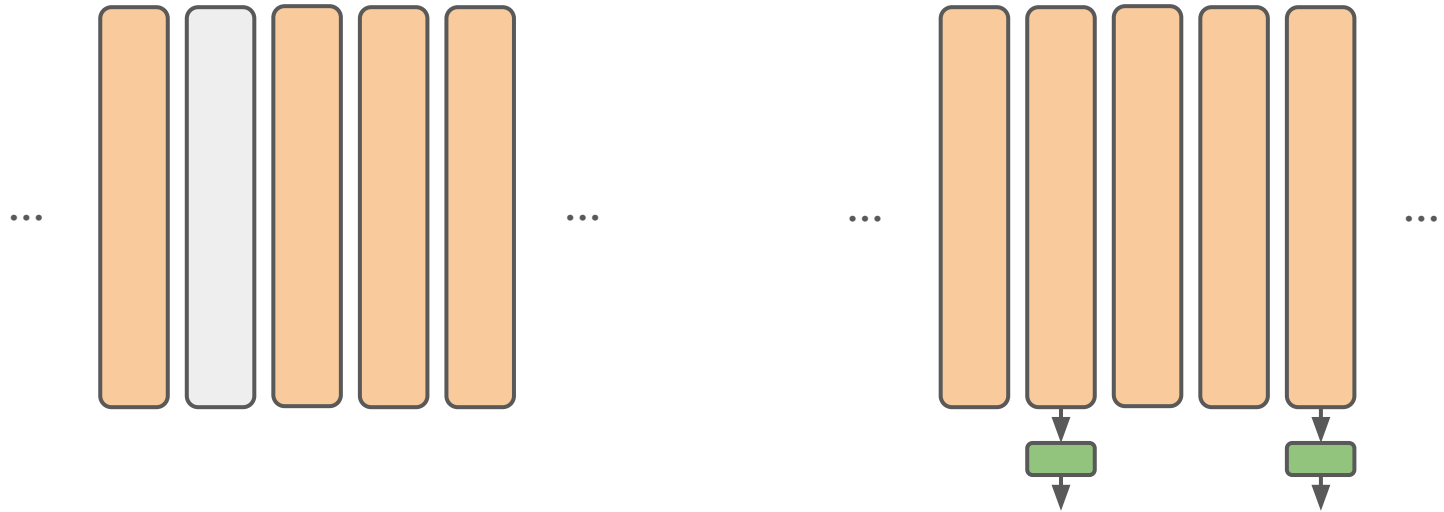
Background

Probing layers



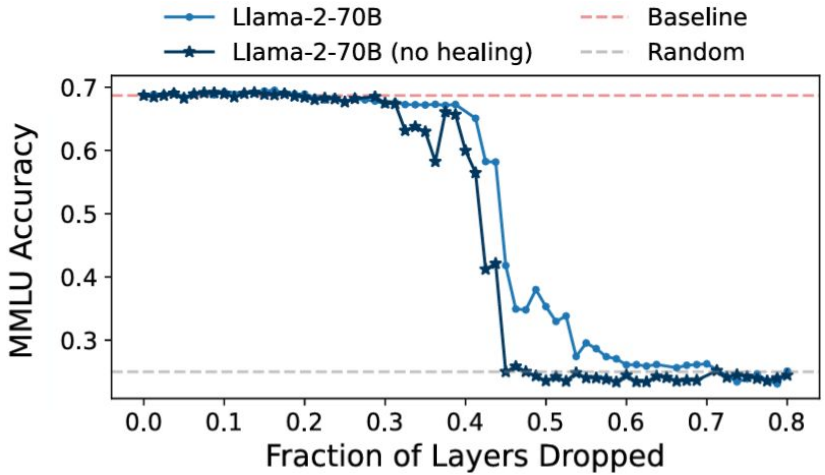
Background

Probing layers



Background

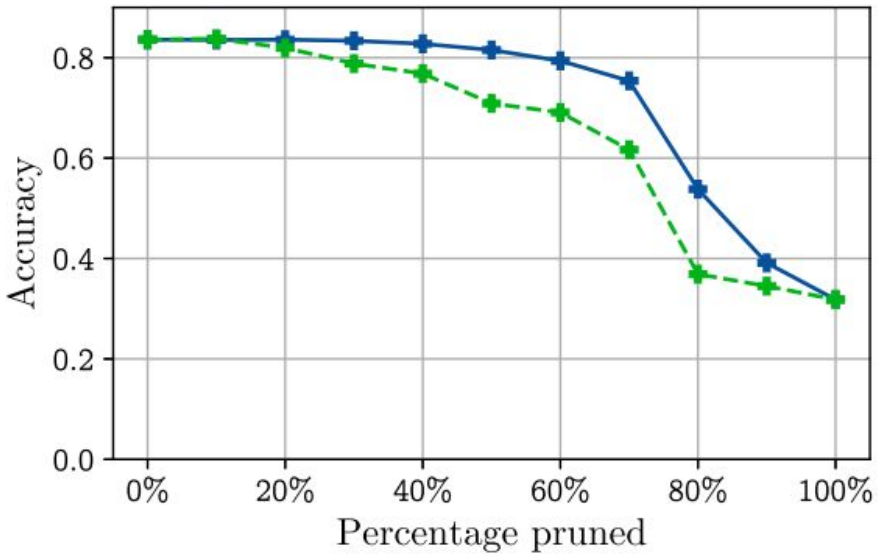
Pruning layers has minimal performance degradation



Gromov, Andrey, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. "The unreasonable ineffectiveness of the deeper layers." arXiv preprint arXiv:2403.17887 (2024).

Background

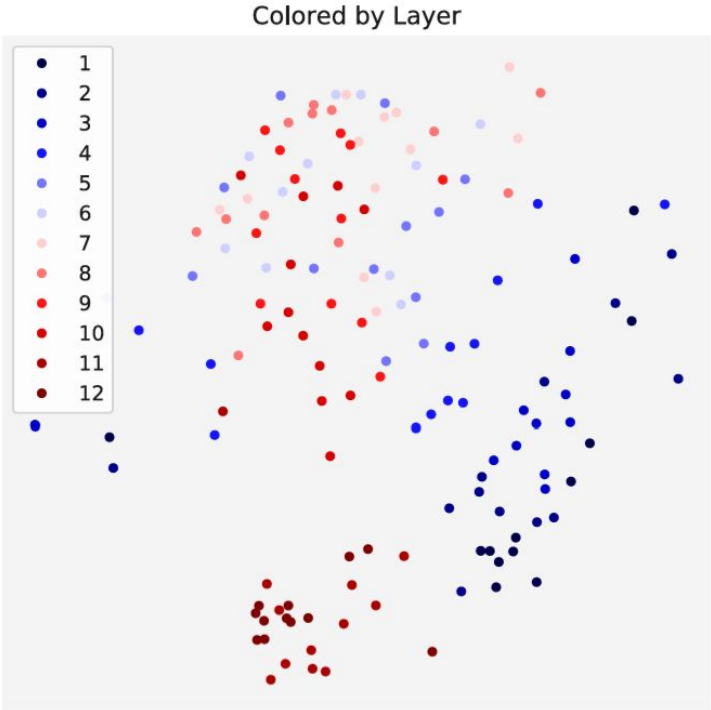
Pruning attention heads



Michel, Paul, Omer Levy, and Graham Neubig. "Are sixteen heads really better than one?." Advances in neural information processing systems 32 (2019).

Background

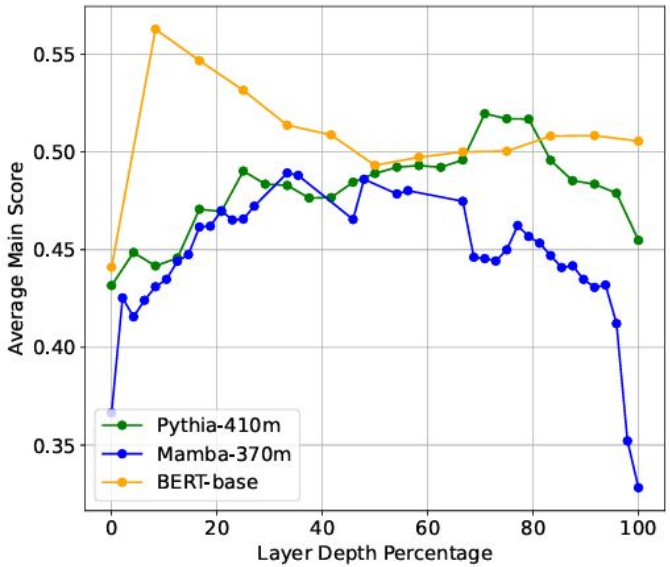
Layer clusters



Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. "What does BERT look at? an analysis of BERT's attention." In Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, pp. 276-286. 2019.

Background

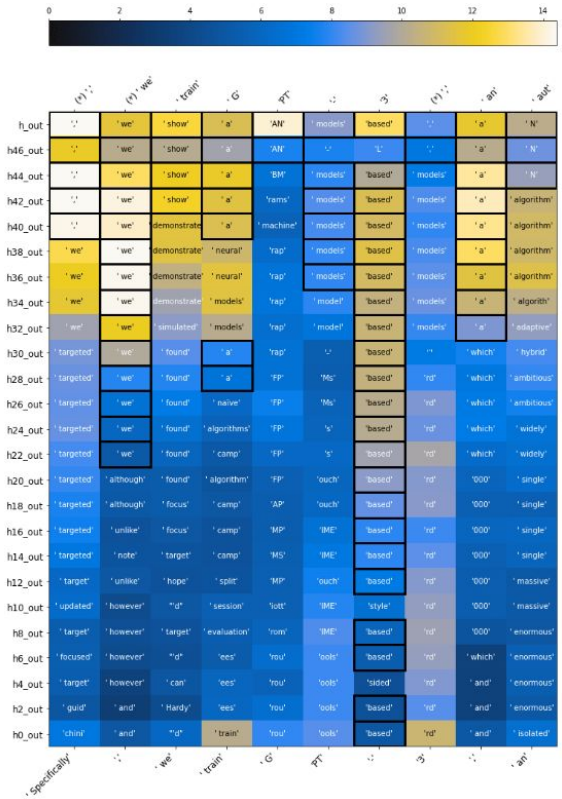
Intermediate representations outperform final ones



Skean, Oscar, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. "Layer by layer: Uncovering hidden representations in language models." arXiv preprint arXiv:2502.02013 (2025).

Background

LogitLens



Nostalgebraist. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020

Do Language Models Use Their Depth Efficiently?

Róbert Csordás **Christopher D. Manning** **Christopher Potts**
Stanford University, Stanford, CA, USA
rcsordas@cs.stanford.edu {manning,cgpotts}@stanford.edu

Background

The residual

$$a_\ell = \text{SelfAttention}_\ell(\text{Norm}(h_\ell))$$

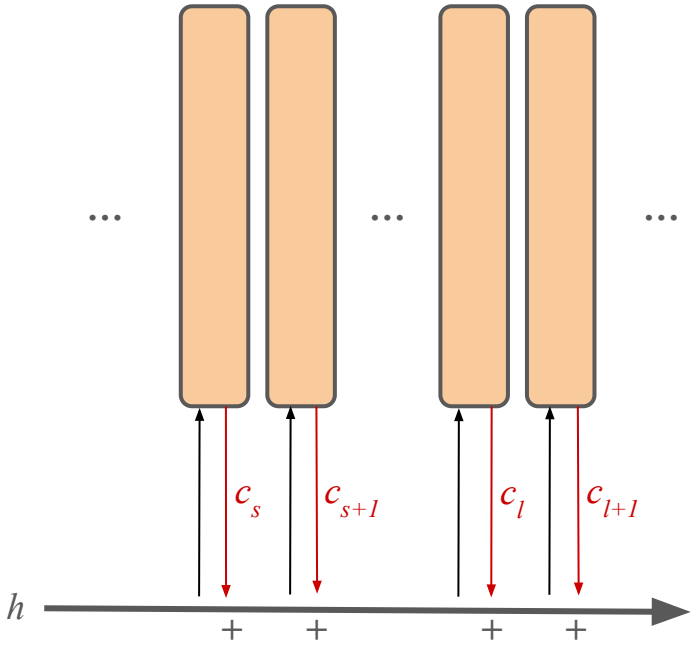
$$\hat{h}_\ell = h_\ell + a_\ell$$

$$m_\ell = \text{MLP}_\ell(\text{Norm}(\hat{h}_\ell))$$

$$h_{\ell+1} = \hat{h}_\ell + m_\ell$$

Background

The residual

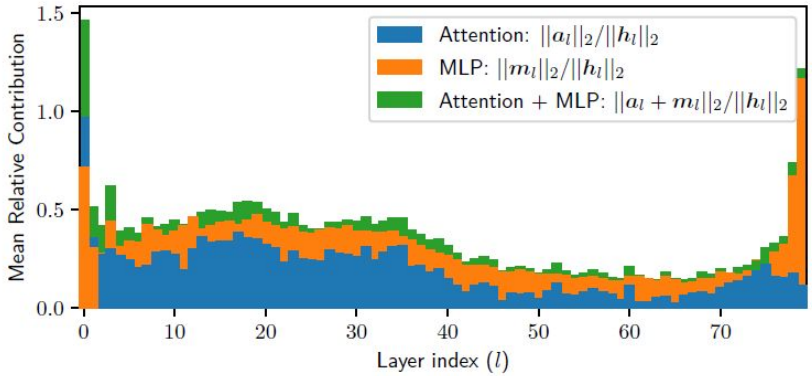


$$c_x = a_x + m_x$$

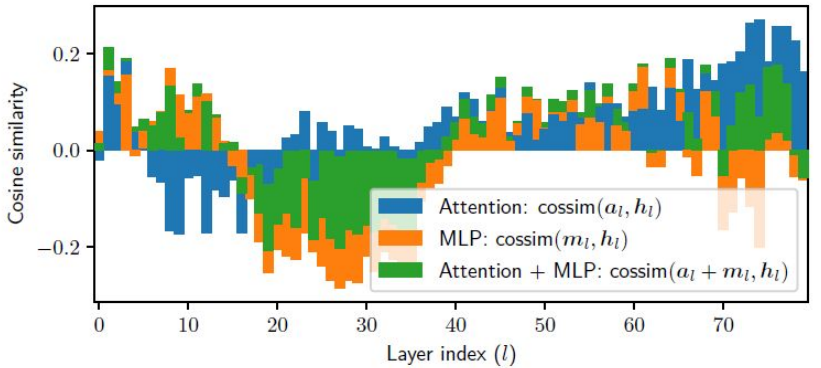
Experiment 1:
How do the layers contribute to the residual stream?

How do the layers contribute to the residual stream?

(Sub)layer contributions



(a) Relative norm of (sub)layer contributions

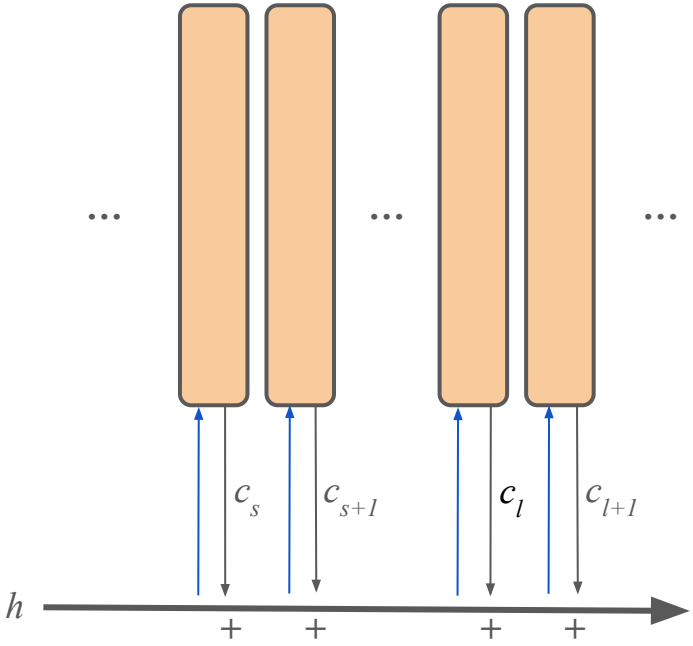


(b) Cossims of (sub)layer contributions and the residual

Experiment 2:
How important is each layer for downstream computation?

How important is each layer for downstream computation?

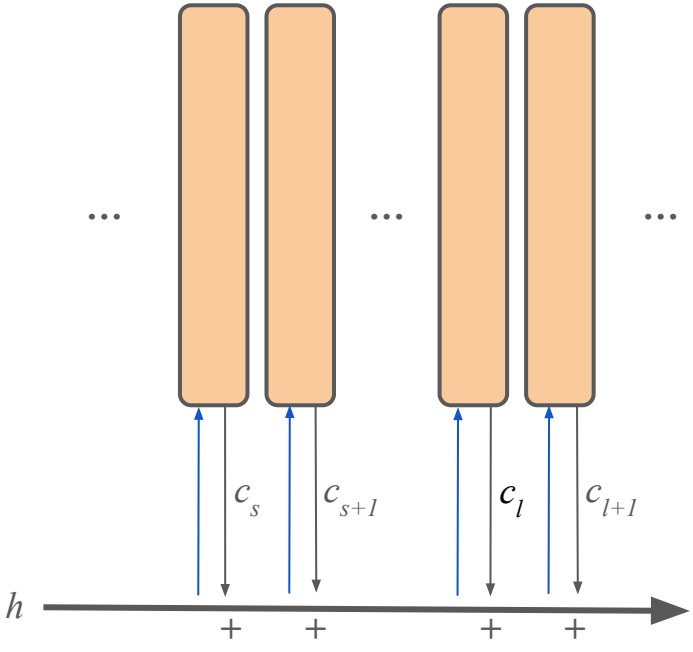
Zeroing layer contributions



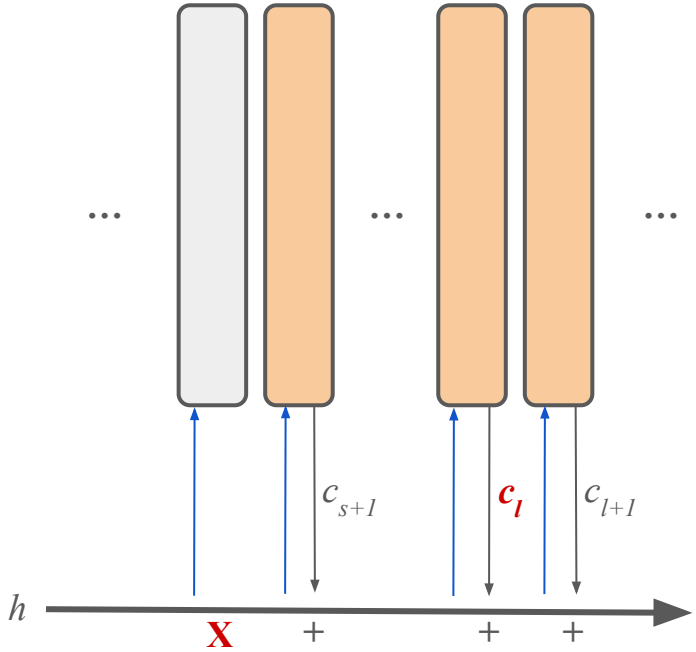
$$c_x = a_x + m_x$$

How important is each layer for downstream computation?

Zeroing layer contributions

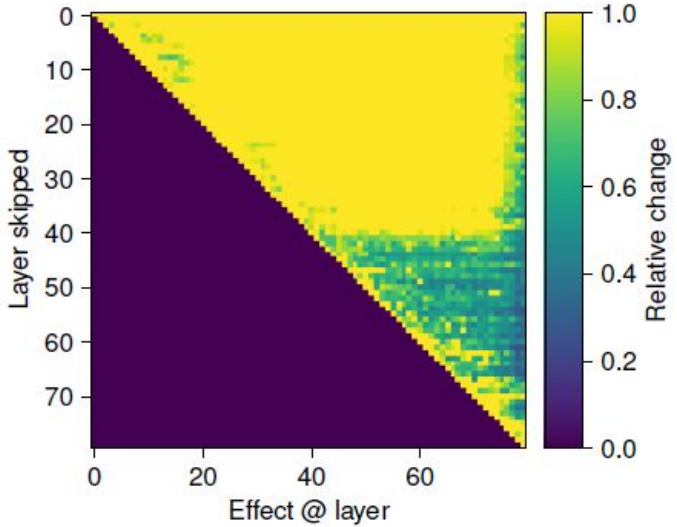


$$c_x = a_x + m_x$$



How important is each layer for downstream computation?

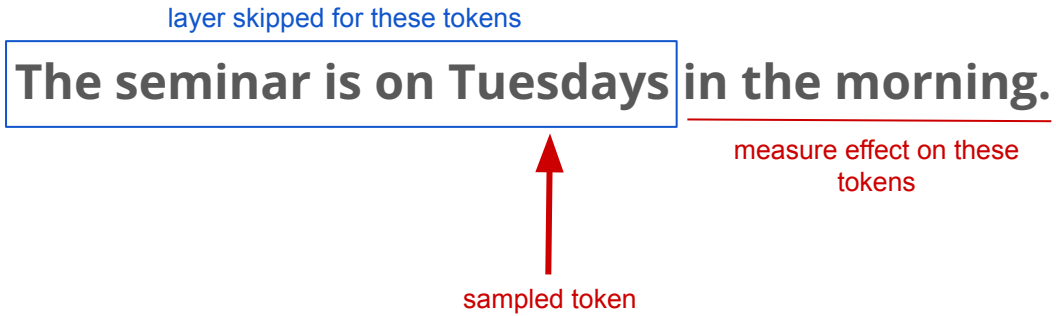
Effect of skipping each layer (current tokens)



(a) Effect of skipping a layer on later layers contributions in the *all* timesteps.

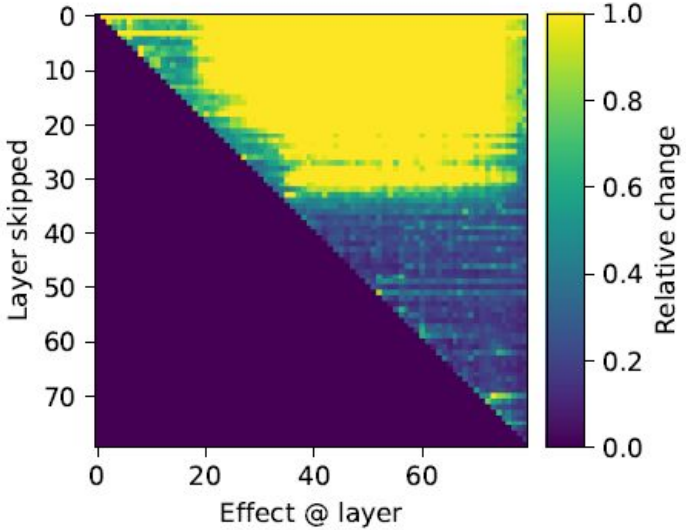
How important is each layer for downstream computation?

Effect of skipping each layer (future tokens)



How important is each layer for downstream computation?

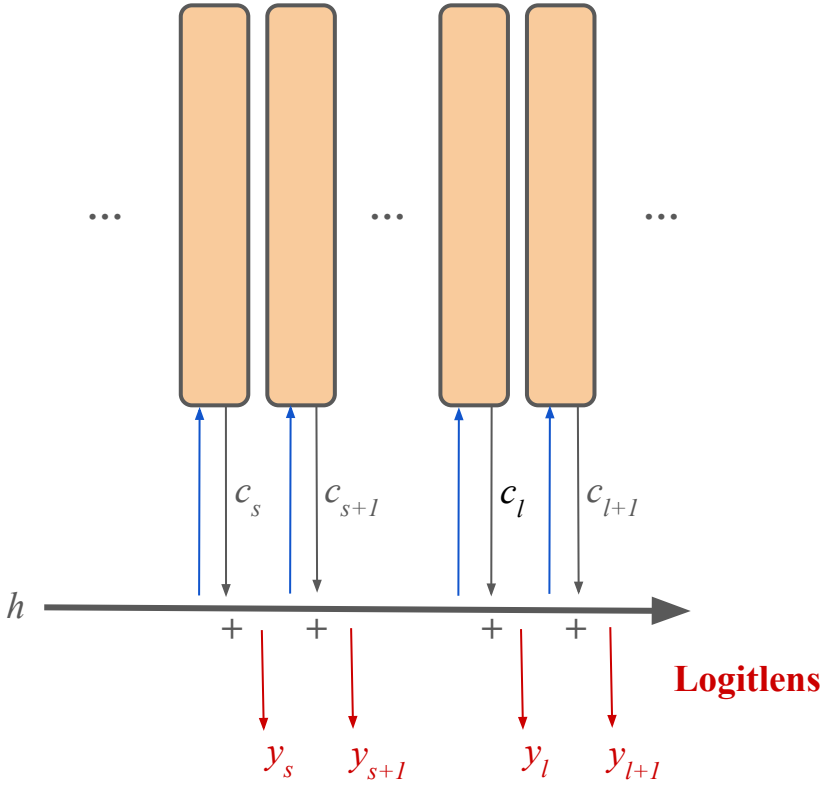
Effect of skipping each layer (future tokens)

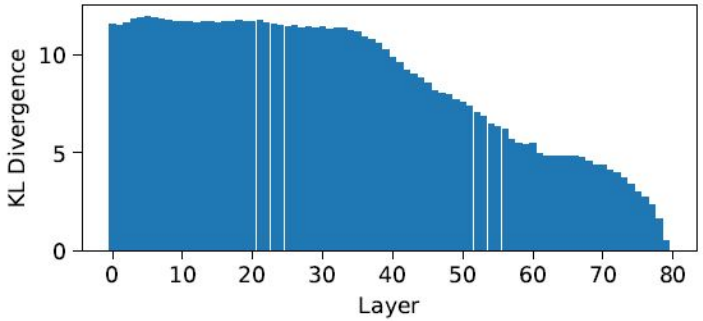


(b) Effect of skipping a layer on later layers' contributions in *future* timesteps.

How important is each layer for downstream computation?

What happens in the second half of the network?



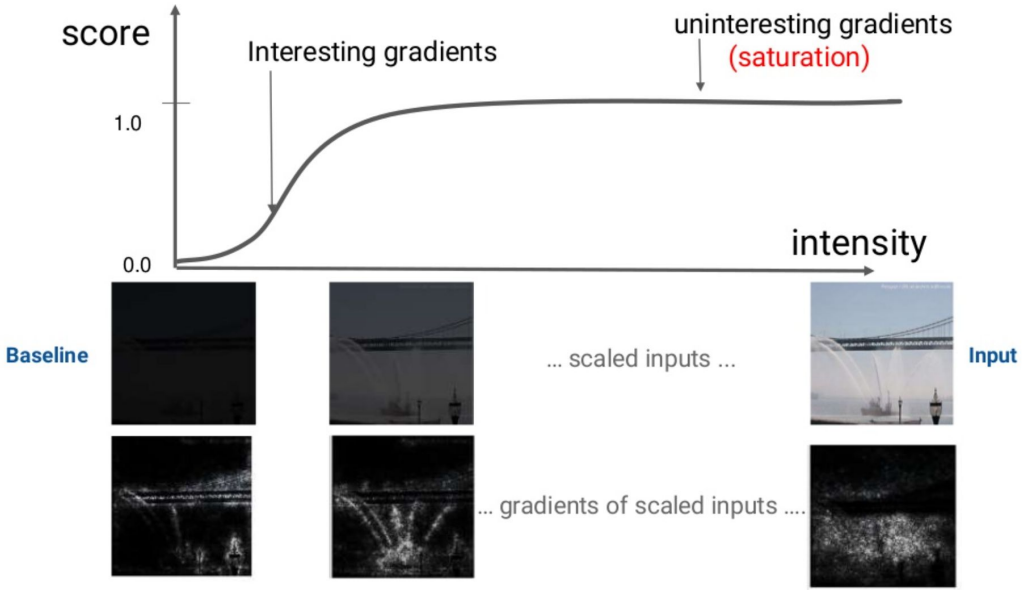


(a) KL divergence between the model's prediction and Logitlens applied to each layer.

Experiment 3:
Do deeper problems use deeper computation?

Background

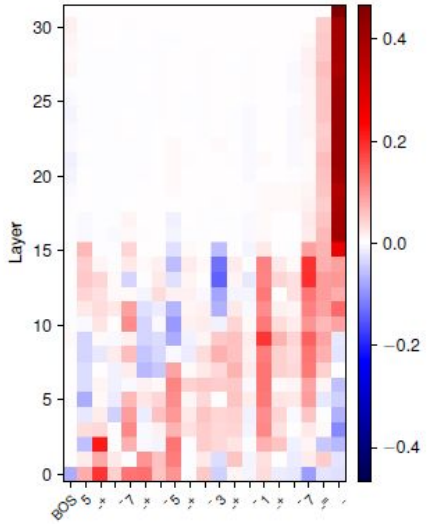
Integrated gradients



Pai. Interpreting Deep Neural Networks using Integrated Gradients https://towardsdatascience.com/interpreting-deep-neural-networks-using-integrated-gradients-f9b8ecdd3c57, 2020

Do deeper problems use deeper computation?

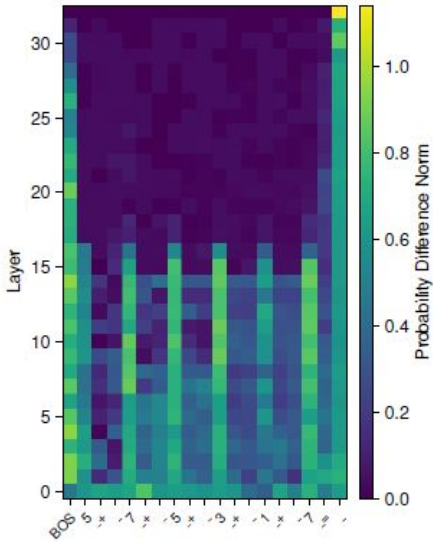
Integrated gradients



(a) Integrated gradients on a math question.

Do deeper problems use deeper computation?

Residual erasure

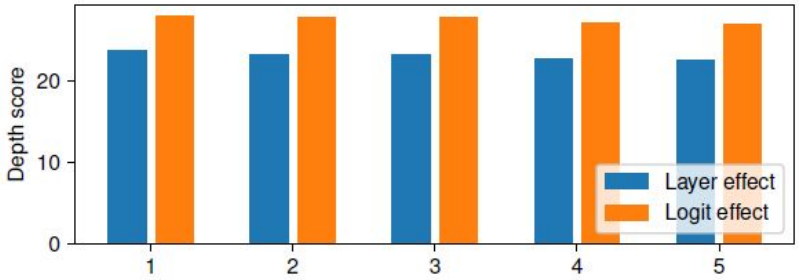


(b) Residual erasure on a math question.

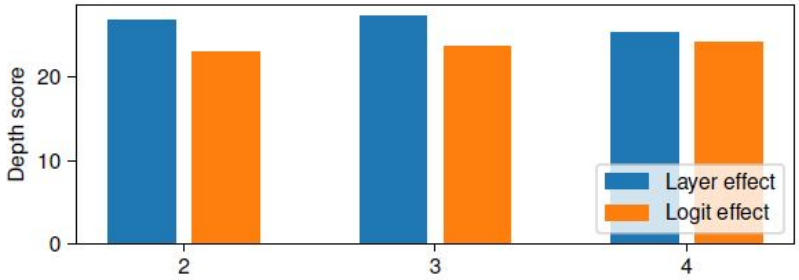
Do deeper problems use deeper computation?

Depth score

$$d = \sum_{l=1}^{L} l \frac{e_l}{\sum_m e_m}$$



(a) Depth score on different difficulty levels of MATH

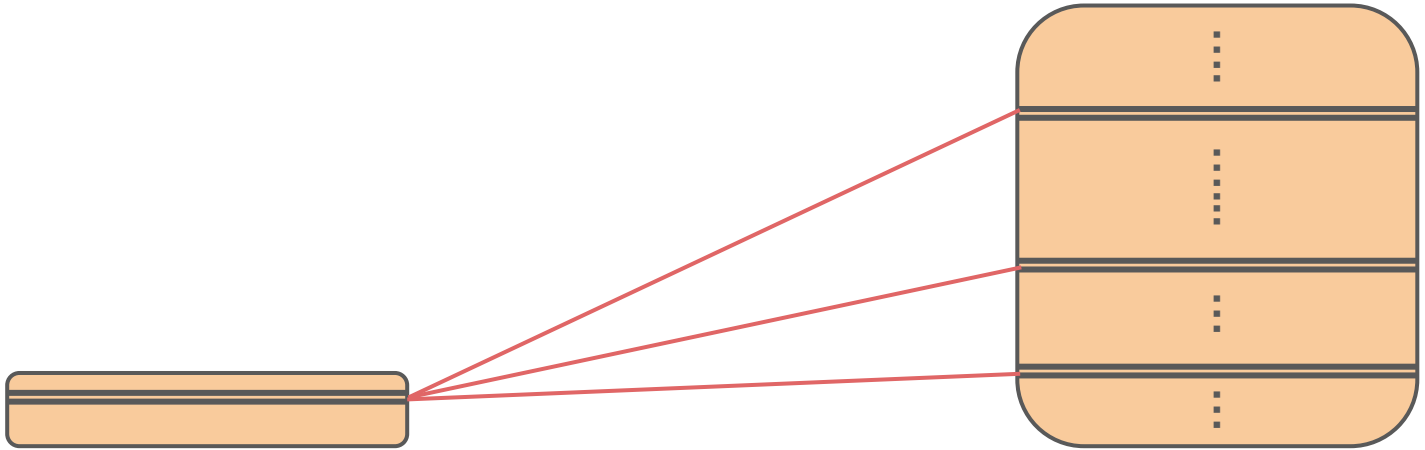


(b) Depth score in function of hops on MQuAKE

Experiment 4:
Do deeper models do novel computation?

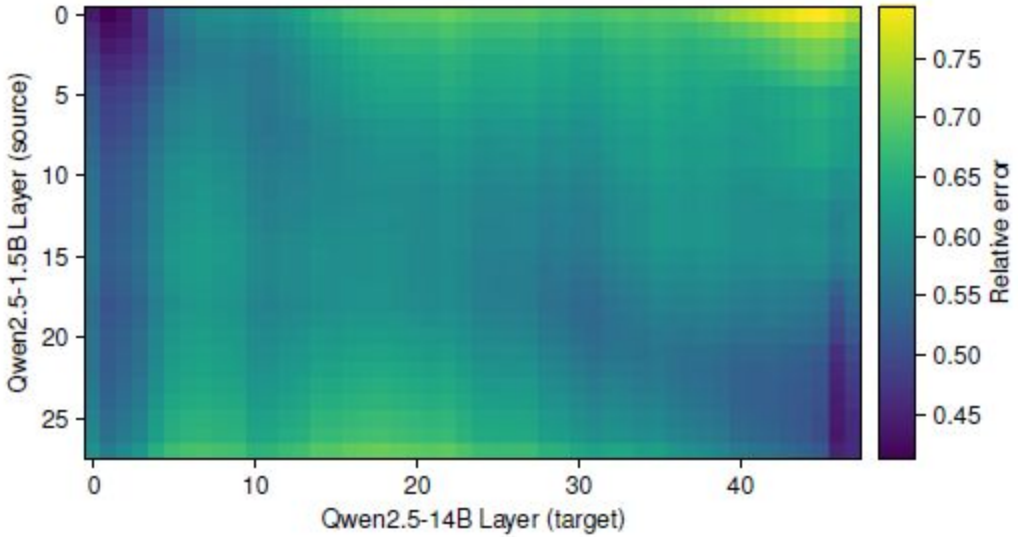
Do deeper models do novel computation?

Mapping representations from shallow model to deep model



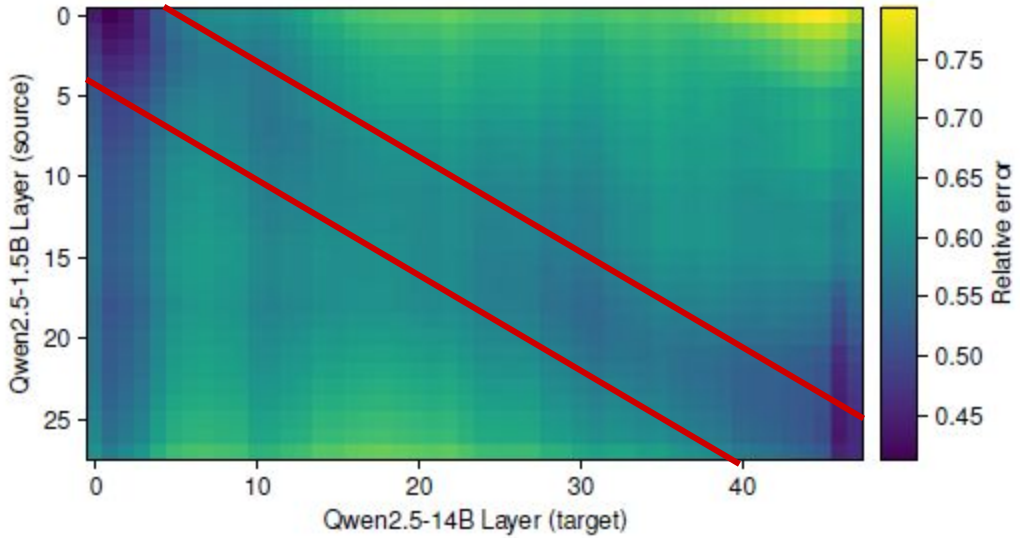
Do deeper models do novel computation?

Deeper models “spread out” computations



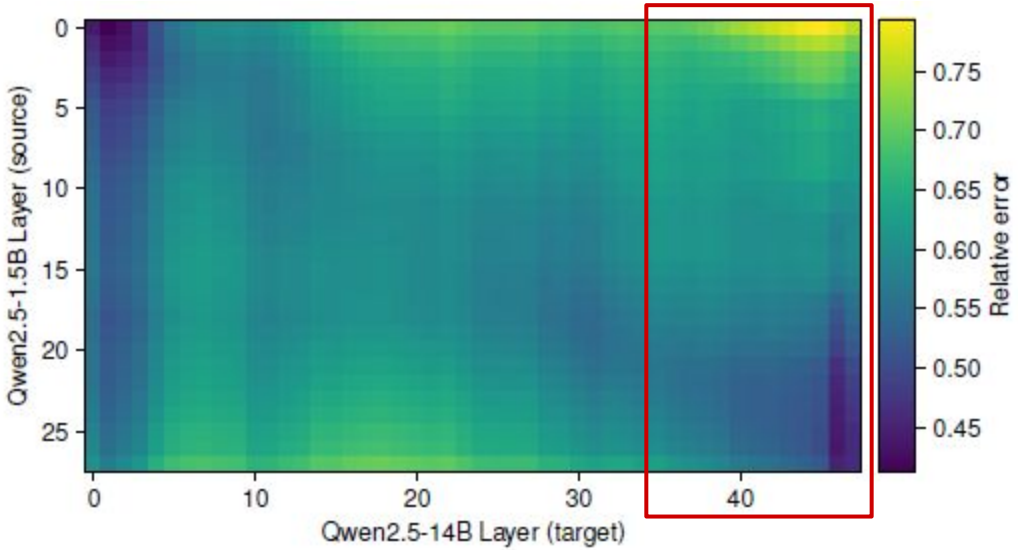
Do deeper models do novel computation?

Deeper models “spread out” computations



Do deeper models do novel computation?

Deeper models “spread out” computations



Evaluation

Strengths and limitations

- **Strengths**
  - Rigorous and diverse experiments
  - Practical implications for depth scaling and architectures
  - Interpretability contribution
- **Limitations**
  - Generalizability of the findings beyond models, tasks, and data
  - Limited mechanistic or causality explanations
  - Focused on language models, reasoning models not tested

Takeaways

Later layers contribute much less computation

Deeper problems ≠ deeper reasoning

Deeper models ≠ deeper reasoning