Paper #840 • NAACL ’24

The Impact of Depth on Compositional Generalization in Transformer Language Models

Jackson Petty[1*], Sjoerd van Steenkiste[2], Ishita Dasgupta[3], Fei Sha[2], Dan Garrette[3], and Tal Linzen[2]

[1] NYU
[2] Google Research
[3] Google DeepMind

* *Work done as a Student Researcher at Google Research*

Up-front Conclusions

Past work: **depth** increases expressive capacity & experimentally improves transformer LM performance on linguistic tasks

We ask: **Are deeper transformer language models more *compositional*, independent of total parameter count?**

Up-front Conclusions

Past work: **depth** increases expressive capacity & experimentally improves transformer LM performance on linguistic tasks

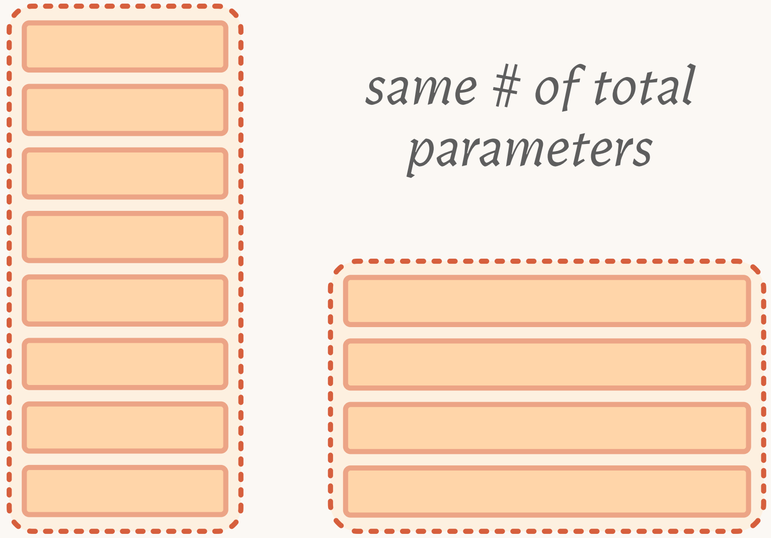
We ask: **Are deeper transformer language models more *compositional*, independent of total parameter count?**

When controlling for the total # of parameters:

1. **deeper models are better** at both language modeling and compositional generalization;

2. depth has **diminishing marginal utility** (versus width) for improving performance on language modeling and compositional generalization;

3. depth has a **linear computational cost** (versus width).



The Details

What did we study, and why did we conclude this?

What is “compositionality”?

Compositionality is the “infinite use of finite means”

(von Humbolt 1836; Chomsky 1965)

Generalize from **known pieces** to **(infinite) novel, well-formed** combinations

Training input (hedgehog is subject)		*Output*
`the hedgehog ate the cake`		*eat*(***agent=hedgehog***, *theme=cake*)
`the hedgehog saw a child`		*see*(***agent=hedgehog***, *theme=child*)
`hedgehogs swim`	→	*swim*(***agent=hedgehog***)

Semantic parsing example from COGS (vf)

What is “compositionality”?

Compositionality is the “infinite use of finite means”

(von Humbolt 1836; Chomsky 1965)

Generalize from **known pieces** to **(infinite) novel, well-formed** combinations

Training input (hedgehog is subject)		*Output*
`the hedgehog ate the cake`		*eat*(***agent=hedgehog***, *theme=cake*)
`the hedgehog saw a child`		*see*(***agent=hedgehog***, *theme=child*)
`hedgehogs swim`	→	*swim*(***agent=hedgehog***)
Generalization (hedgehog is object)		
`the boy loves the hedgehog`		*love*(*agent=boy*, ***theme=hedgehog***)

Semantic parsing example from COGS (vf)

Why do we care about compositionality?

Humans are quite good at generalizing OOD on compositional tasks; current models are less good

Compositional leaners make **more human-like** generalizations, require **less data**, **more robust** to (some) distribution shifts

Compositionality shows up in:

- semantic parsing
- code generation
- natural language understanding
- & more!

Research Question

We ask: **Are deeper models more *compositional*, independent of total parameter count?**

Intuition, informed by theory and experimental results, that *deeper* models are often *better* models

Theory:

- Expressive capacity ~ exponential in depth
- Depth = successive function application

Empirically:

- Reducing depth harms linguistic generalization more than width does

How do we test if depth is helpful?

We ask: **Are deeper models more *compositional*, independent of total parameter count?**

Depth (# of layers) *and size* ***(total # of parameters)*** *are often confounded!*

Adding (independent) layers ⟶ adding parameters

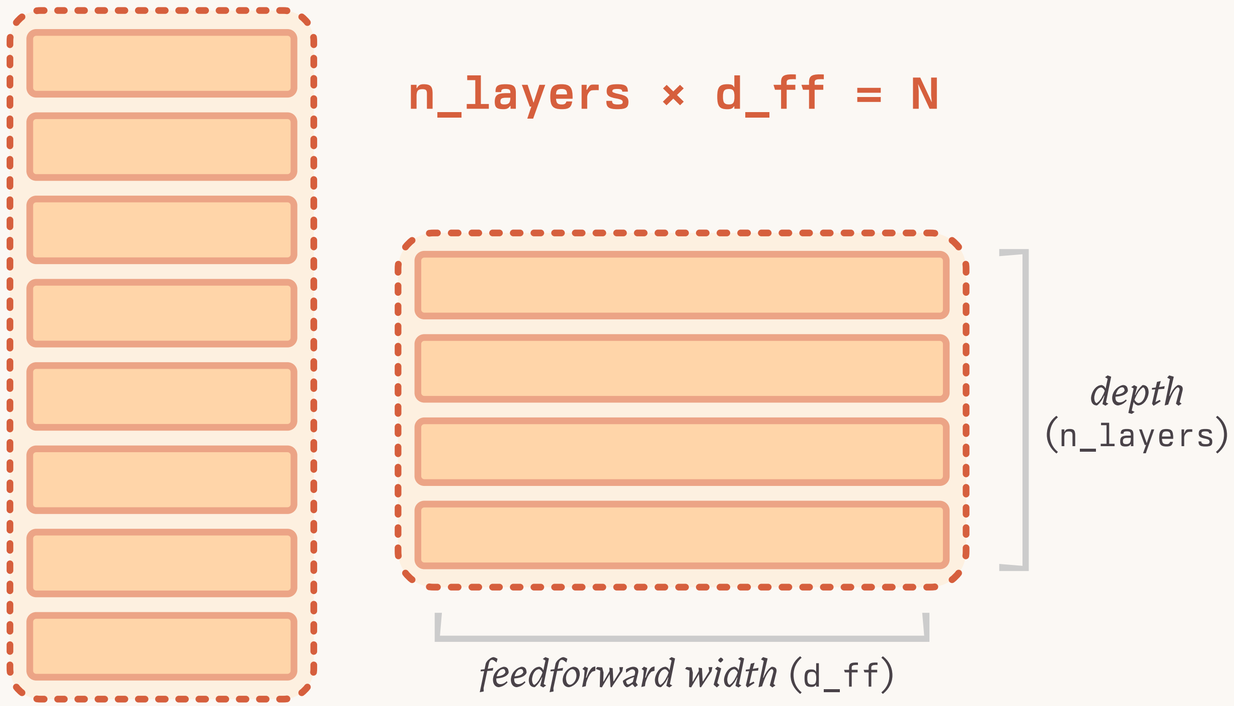
Many things get better when models have more parameters!

To test our hypothesis, we must disentangle depth and size

Separating *depth* from *size*

We construct three size classes of equally-sized transformer models

Within each size class, models have different shapes



Within each size class, all models have the same # of total parameters

Experimental Setup

Experimental Setup

1. pretrain each model with a causal language modeling objective

dataset	*type*	*metric*
C4-en	language modeling	validation loss

Experimental Setup

1. pretrain each model with a causal language modeling objective
2. finetune on datasets for compositional generalization

dataset	*type*	*metric*
C4-en	language modeling	validation loss
COGS	semantic parsing	full-sequence generalization accuracy
COGS (variable free)	semantic parsing	
GeoQuery	SQL generation	
English Passivization (EP)	natural language transformation	

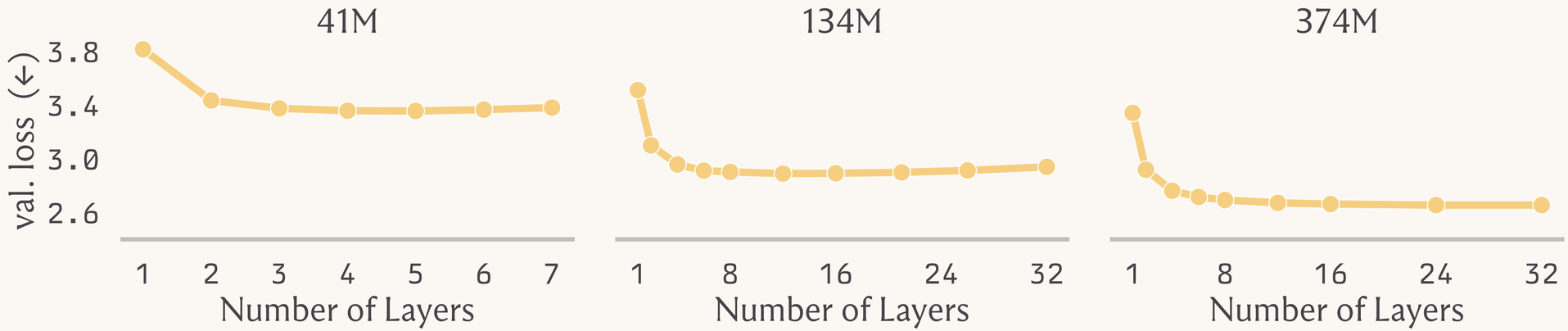
performance vs depth *within each size class* shows impact of depth

Result 1: Language Modeling

Deeper is better, but diminishingly so

As models get deeper, validation perplexity on C4-en goes down

The ***marginal utility*** of depth diminishes very quickly!
1-layer vs 2-layers is much bigger than *12-layers vs 16-layers*

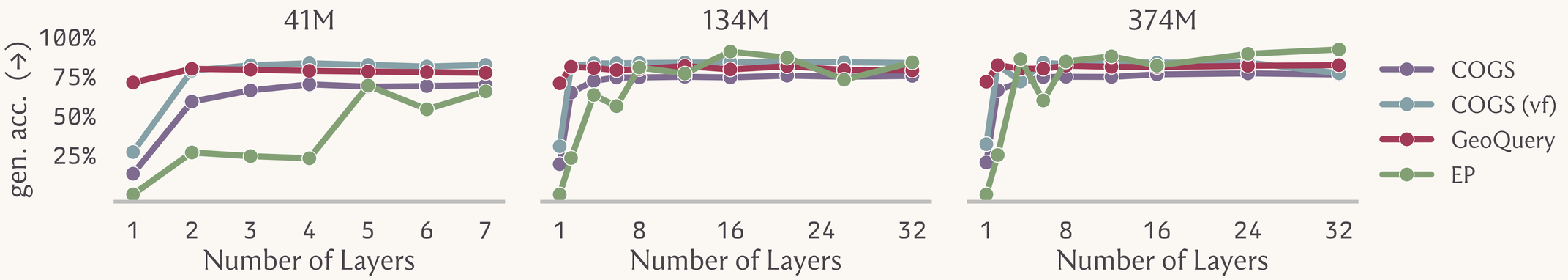


Result 2: Compositional Generalization

Again, deeper is better but diminishingly so

Across 4 datasets, we see similar trends:

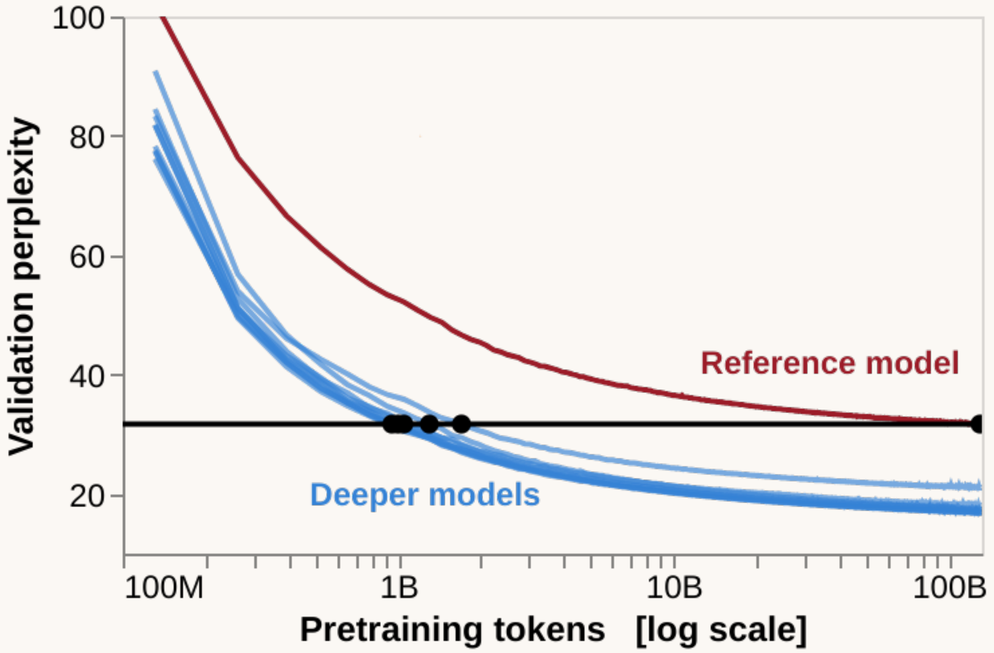
- Very shallow models perform quite poorly
- Once models are deep enough, extra depth isn't useful
- Even deepest models don't get 100% generalization accuracy

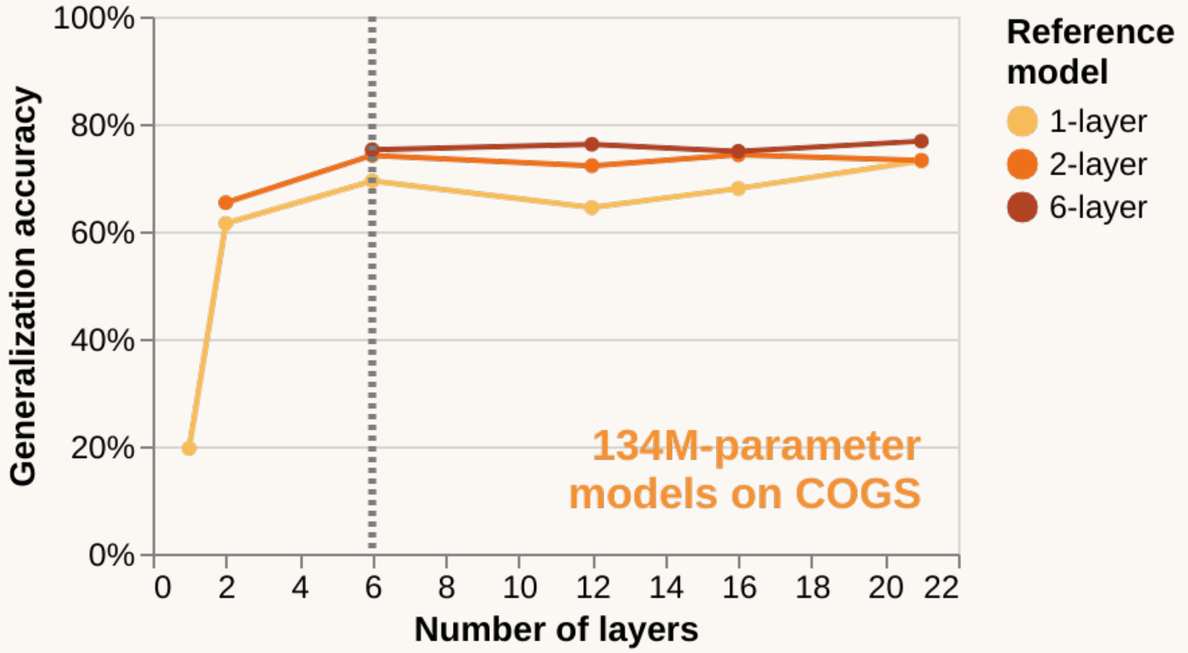


LM & CG results are *independent*

A reasonable worry: what if deeper models are better at compositional generalization *because* they're better at language modeling?

However, correcting for pre-training loss by sampling earlier checkpoints shows that the "depth effect" for compositionality holds independently

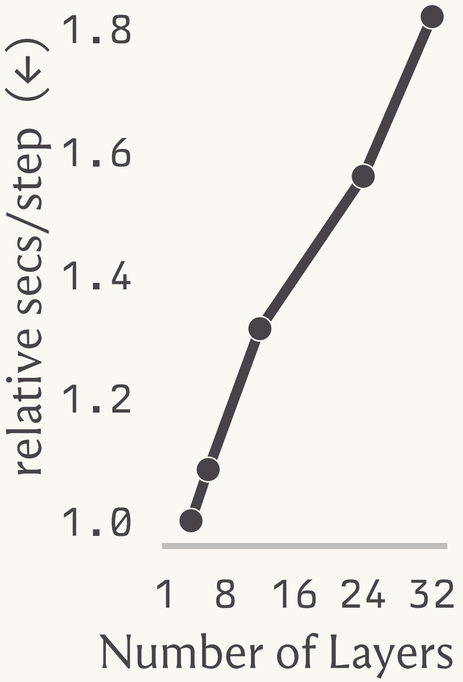


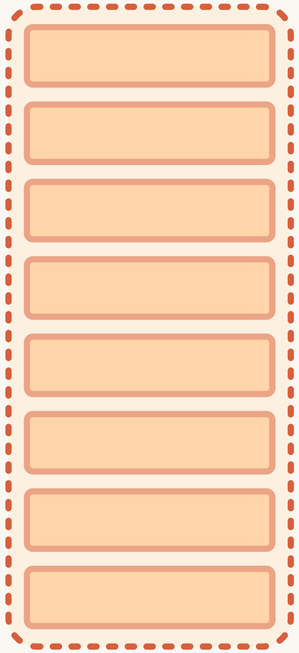


The compute cost of depth is *linear*

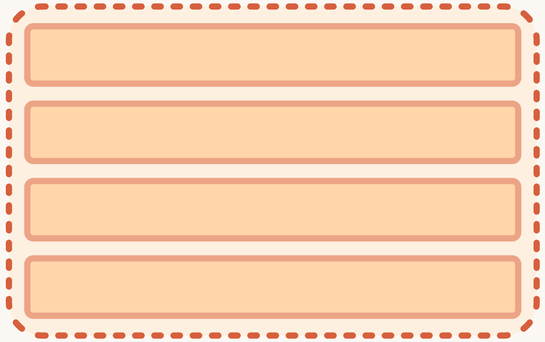
Not all FLOPs are equal: depth introduces sequential dependencies in computation, while width is freely parallelizable

Twice as slow doesn't mean twice as good!





Computation at layer *n* depends on layer *n-1*



What do these results mean for…

…engineers optimizing both performance & cost?

"Deep enough" for our experiments is ~6–8 layers

Scenario #1 (Fixed data volume, minimize time)

"Deep enough" model can train faster than a deeper model, have very close-to-optimal performance, and is cheaper at inference time

Scenario #2 (Fixed time, maximize data volume)

"Deep enough" model can train on *more data* in the same amount of time, possibly exceeding performance of a deeper-but-slower model, and is still cheaper at inference time

What do these results mean for…

…researchers trying to make models more compositional?

Theoretical expressive capacity may not be reflected in everyday tasks.

Deeper models are still exponentially more expressive, but that capacity isn't reflected in LM loss or current compositionality tasks

Compositionality is still hard for all depths & sizes.

Performance saturates but doesn't hit ceiling for our compositionality benchmarks $\longrightarrow$ scale & depth are not all you need here!

Paper #840 • NAACL ’24

The Impact of Depth on Compositional Generalization in Transformer Language Models

Jackson Petty[1*], Sjoerd van Steenkiste[2], Ishita Dasgupta[3], Fei Sha[2], Dan Garrette[3], and Tal Linzen[2]

[1] NYU
[2] Google Research
[3] Google DeepMind

* *Work done as a Student Researcher at Google Research*