BlackBoxNLP @ EMNLP '24

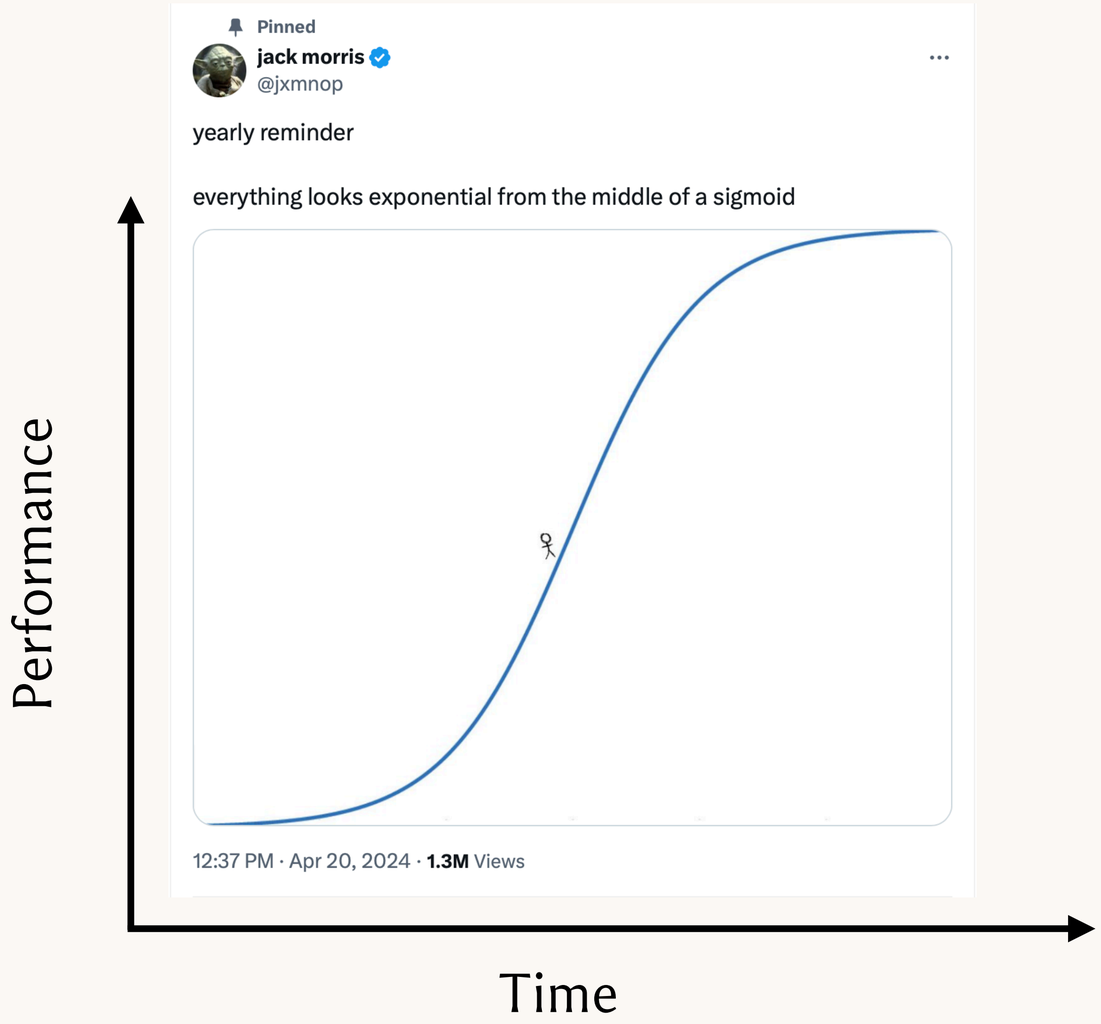
How does code pretraining affect LLM task performance?

Jackson Petty[1*], Sjoerd van Steenkiste[2], and Tal Linzen[2]

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

[1] NYU [2] Google Research

A brief history of LLM research

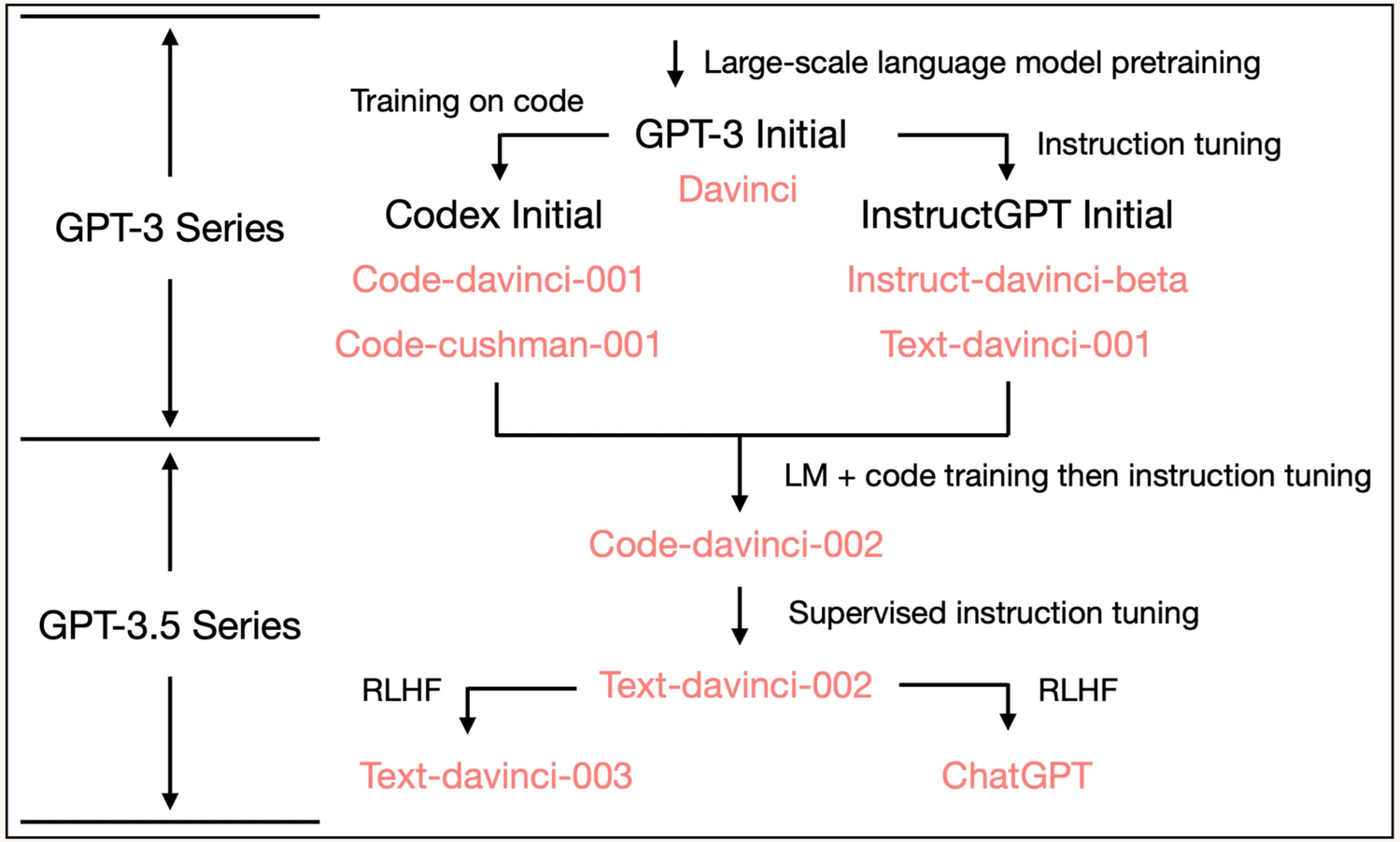


Many things get better with time

It's not always clear why, in hindsight

Historically, adding in code → many improvements in LM performance

What made GPT-3.5 so good?



from *Fu & Khot (2022)*

- Instruction following
- Zero/few-show generalization
- "Reasoning"/Chain-of-Thought

Why might code help LLMs?

Intrinsically:

- Code instantiates many compositional patterns (variable reuse, positional argument structure, literal function composition)
- High quality, low entropy
- Maybe “programming” captures something fundamental:

> Once a programmer knows what to build, the act of writing code can be thought of as (1) breaking a problem down into simpler problems, and (2) mapping those simple problems to existing code (libraries, APIs, or functions) that already exist. The latter activity is probably the least fun part of programming (and the highest barrier to entry), and it’s where OpenAI Codex excels most.

How to measure the impact of code?

Hypothesis: training on code will improve LLM performance on downstream tasks (more than the alternative)

We must control for a lot to test this!

- Model size, dataset size, training hyperparameters…
- Makes observational studies of existing models difficult:
  - Need access to paired models
  - Need to know how models were trained

What is *compositional generalization*?

Compositionality is the "infinite use of finite means"

(von Humboldt 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 *compositional generalization?*

Compositionality is the "infinite use of finite means"

(von Humboldt 1836; Chomsky 1965)

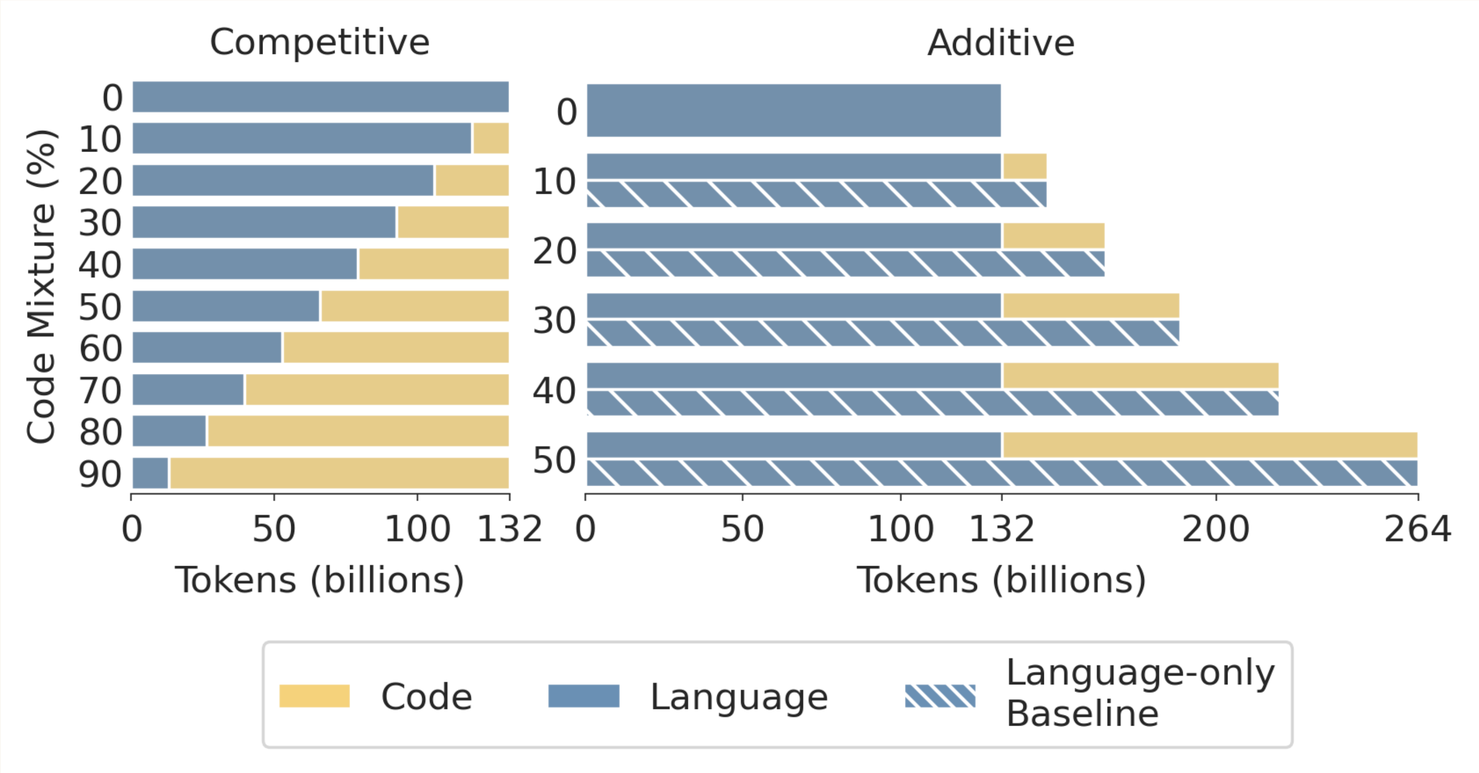
Generalize from **known pieces** to **(infinite) novel, well-formed** combinations

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)

Experimental setup

1. Pretrain (decoder-only) LLMs on mixtures of code (GitHub) and language (C4) in two settings



Experimental setup

1. Pretrain (decoder-only) LLMs on mixtures of code (GitHub) and language (C4) in two settings

2. Finetune each pretrained LLM on compositional-generalization datasets, measure generalization performance

COGS	x:	A hedgehog ate the cake .
	y:	$^*\text{cake}(x_4)$; $\text{hedgehog}(x_1)$ AND $\text{eat.agent}(x_2, x_1)$ AND $\text{eat.theme}(x_2, x_4)$
COGS-vf	x:	A hedgehog ate the cake on the bed .
	y:	$\text{eat}(\text{agent} = \text{hedgehog}, \text{theme} = {}^*\text{cake}(\text{nmod.on} = {}^*\text{bed}))$
English Passivization	x:	our vultures admired her walrus above some zebra .
	y:	her walrus above some zebra was admired by our vultures .

Experimental setup

1. Pretrain (decoder-only) LLMs on mixtures of code (GitHub) and language (C4) in two settings

2. Finetune each pretrained LLM on compositional-generalization datasets, measure generalization performance

3. Evaluate each pretrained LLM on BigBench tasks

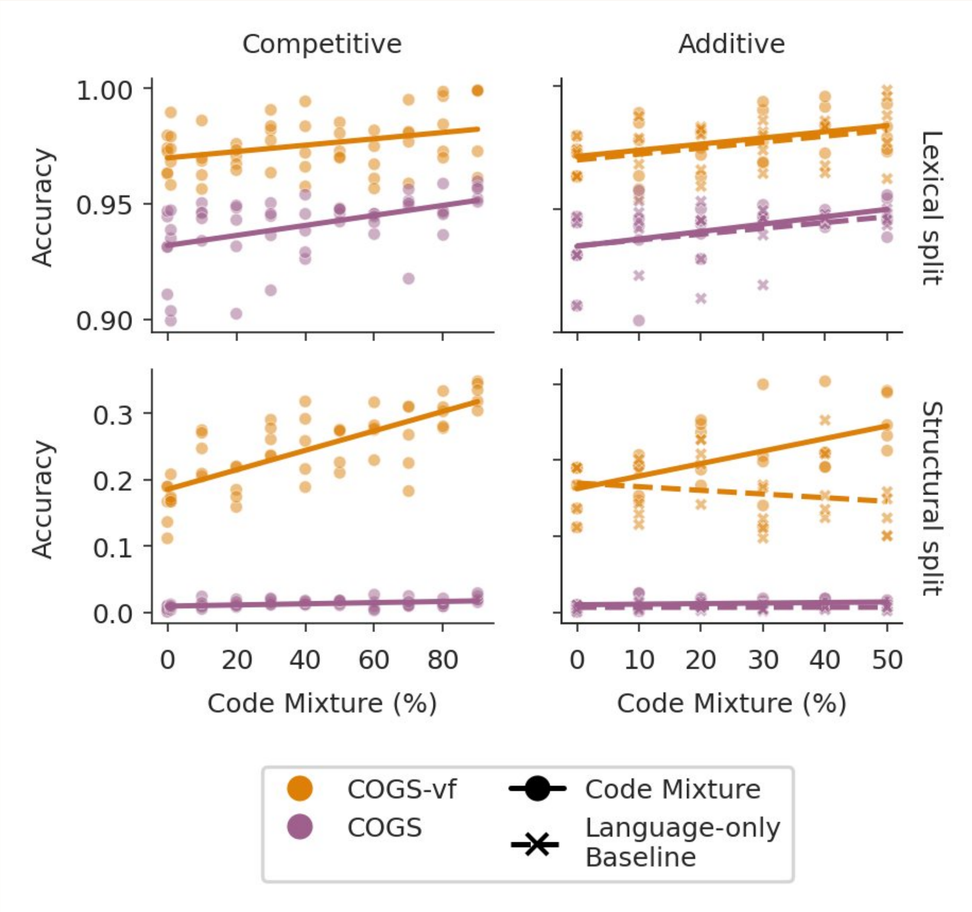
```
What is 42812 times 51570?

(a) 8127851667
(b) 9415523281
(c) 47549647
(d) 854486182733
(e) banana
(f) house
(g) 2207814840
```

```
What is the common morpheme among
these words: biology, antibiotic,
symbiosis, amphibian?

(a) disease
(b) life
(c) study
(d) animal
```

Result #1: Semantic parsing



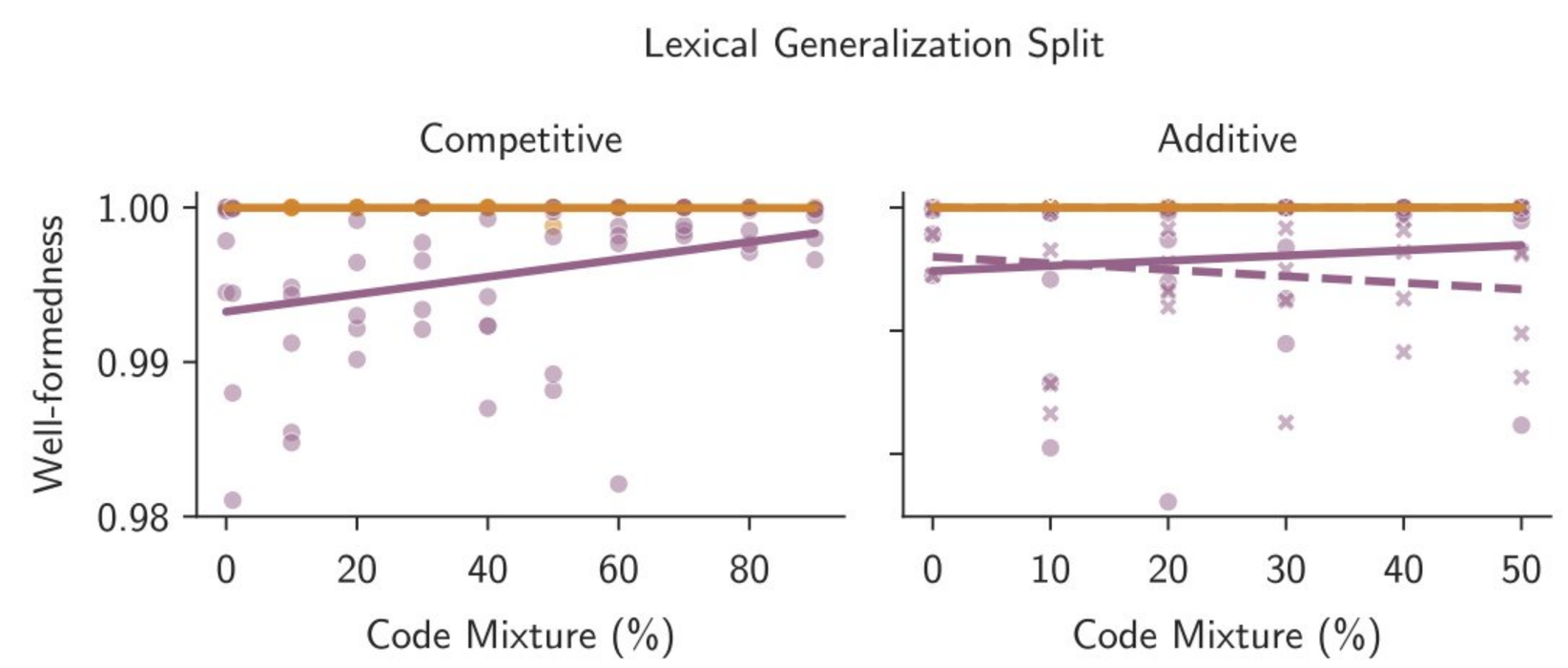
Code-pretraining helps on semantic parsing tasks!

Most noticeable in the “hard” split of COGS-vf

Doesn’t move the needle on the “hard” split of COGS

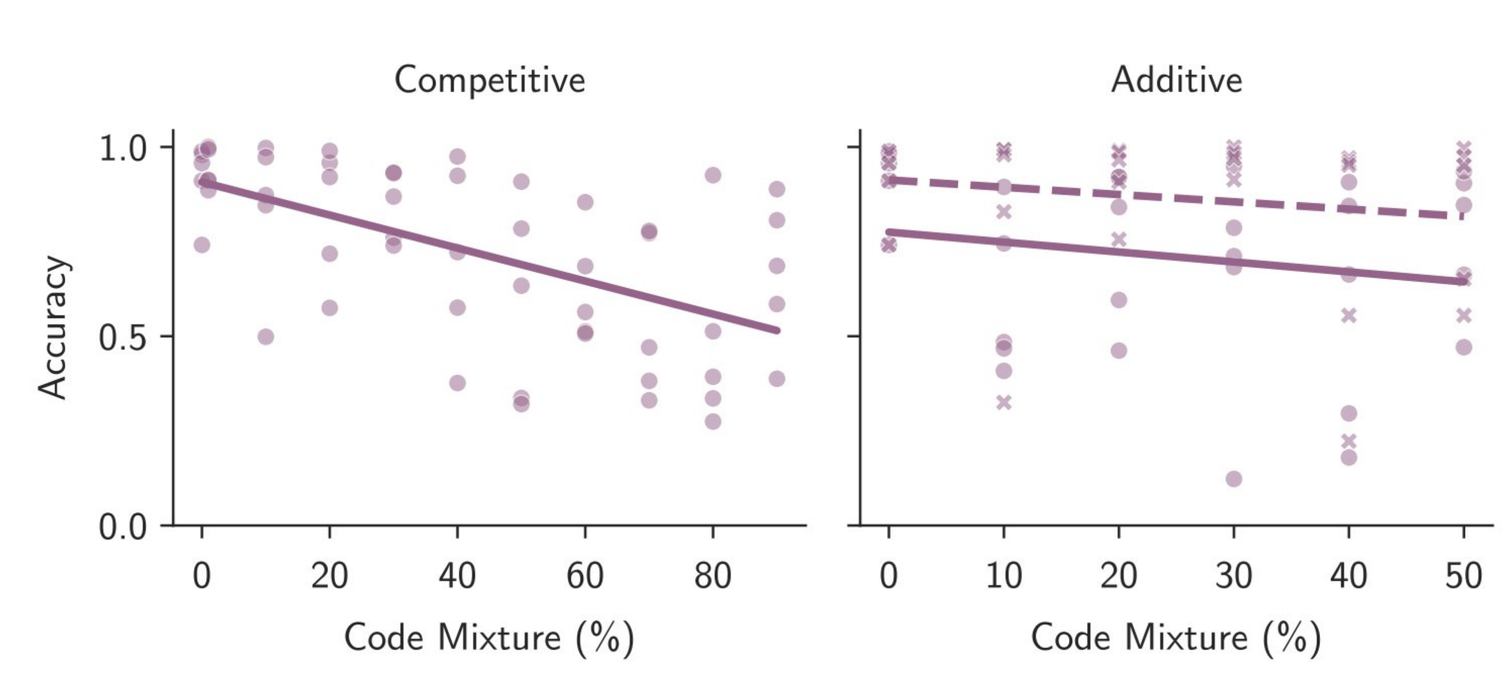
Result #2: Well-formedness

Code actually helps with distributional compositionality, not (just) with well-formedness!



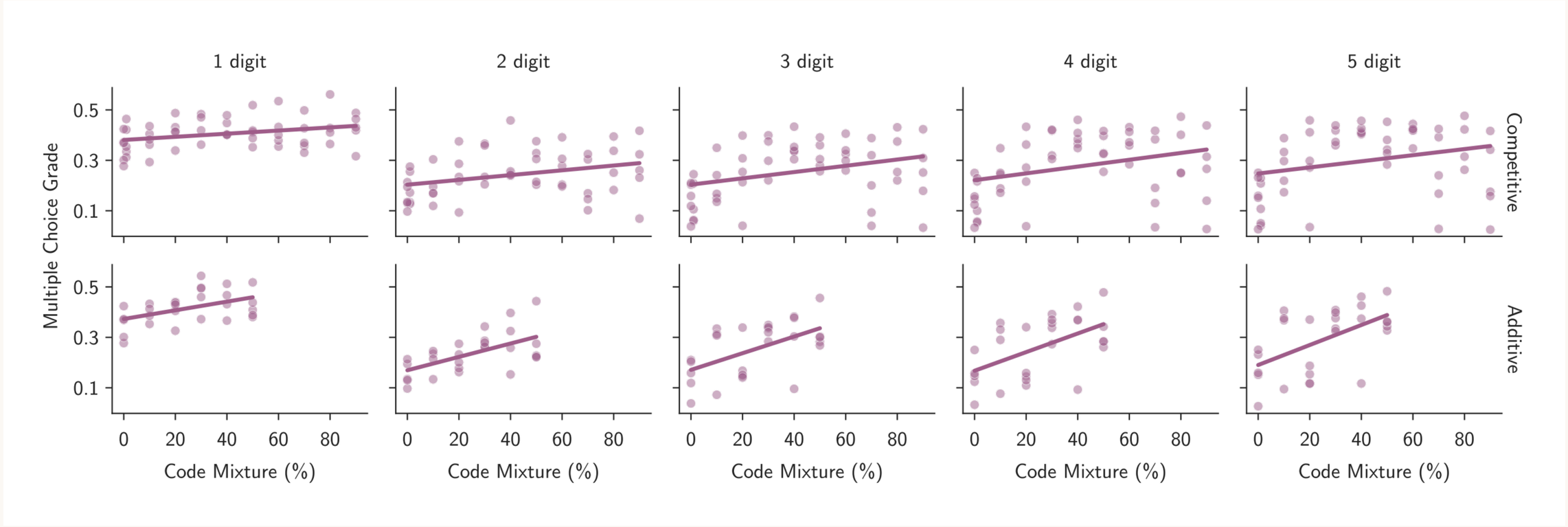
Result #3: NL compositionality

Code harms performance on “natural-language only” compositional generalization (English Passivization)



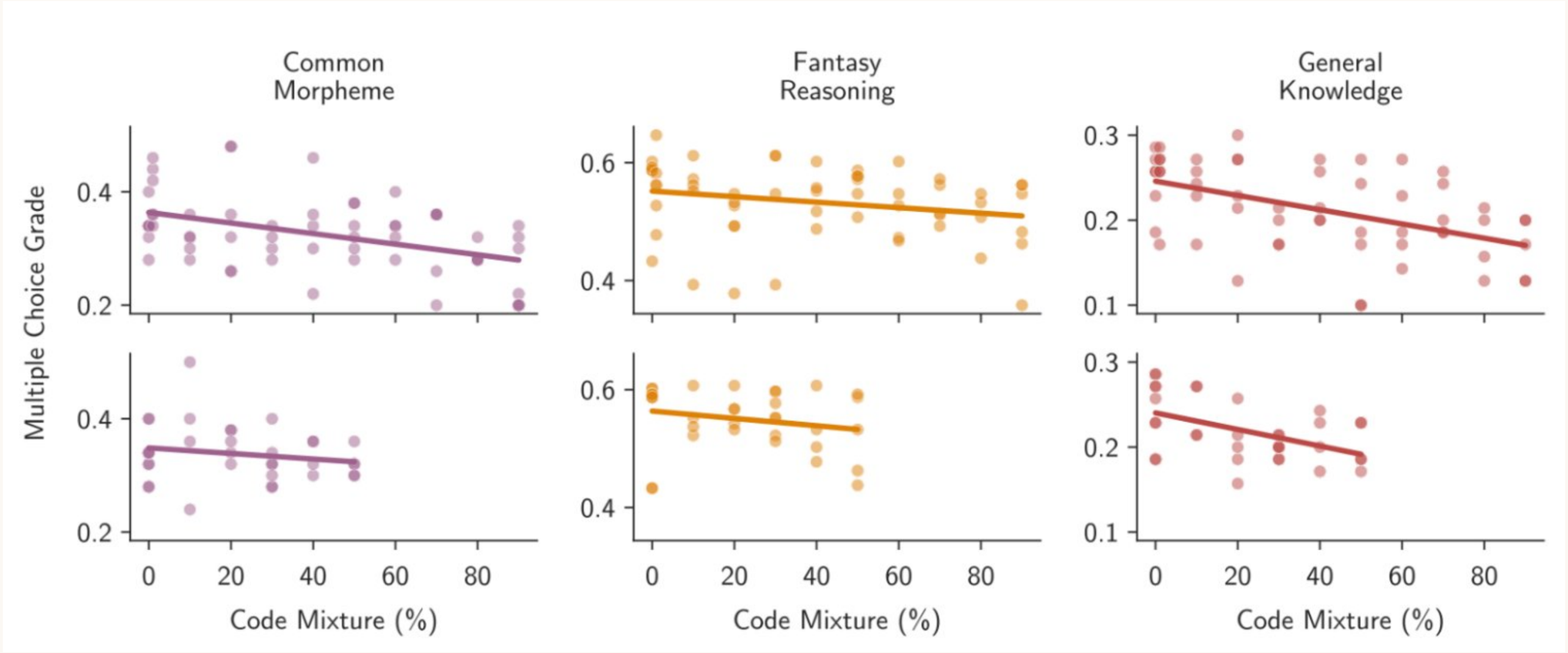
Result #4: Arithmetic

Code improves accuracy on multi-digit arithmetic

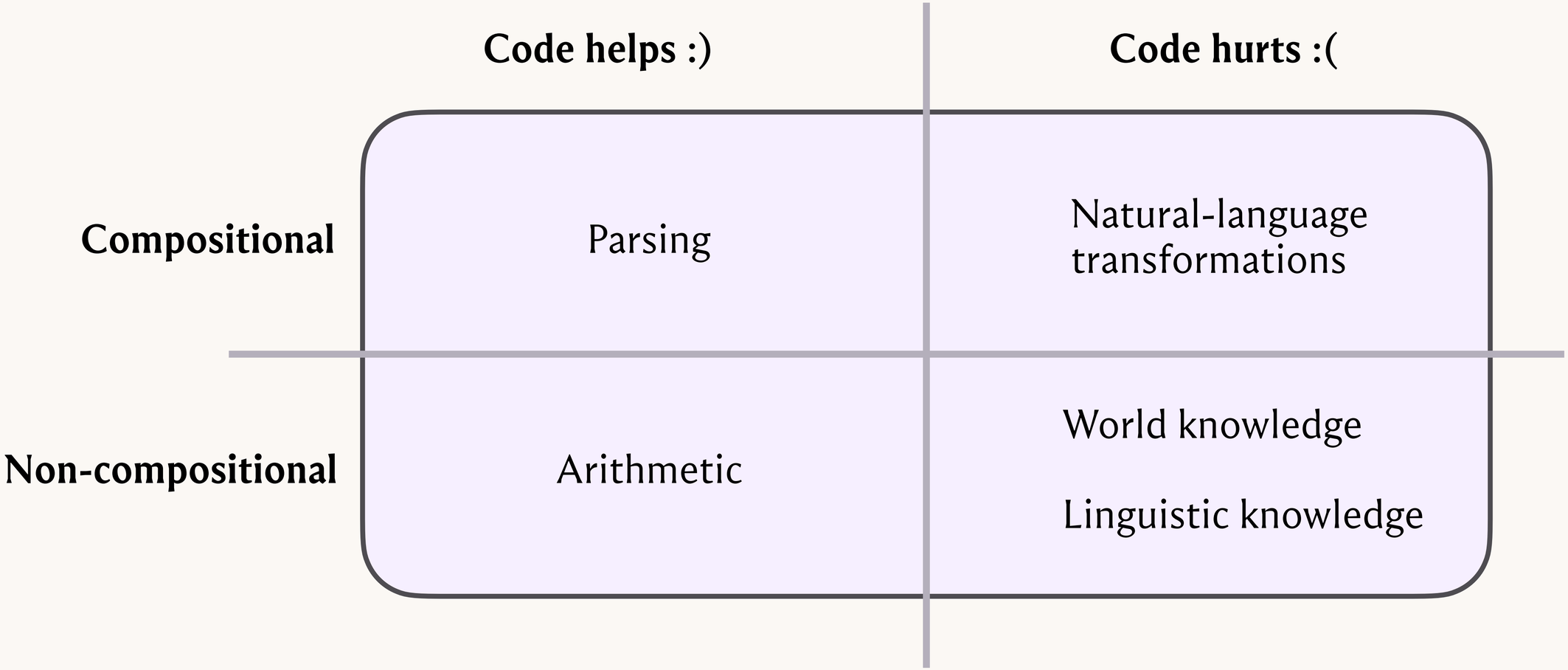


Result #5: “Knowledge”

Code hurts on BB tasks that require linguistic- or world-knowledge



Summary


Code helps :)
Code hurts :(
Compositional
Parsing
Natural-language transformations
Non-compositional
Arithmetic
World knowledge
Linguistic knowledge

Thanks!

Come say “hi” at EMNLP :)

Appendix A: COGS Splits

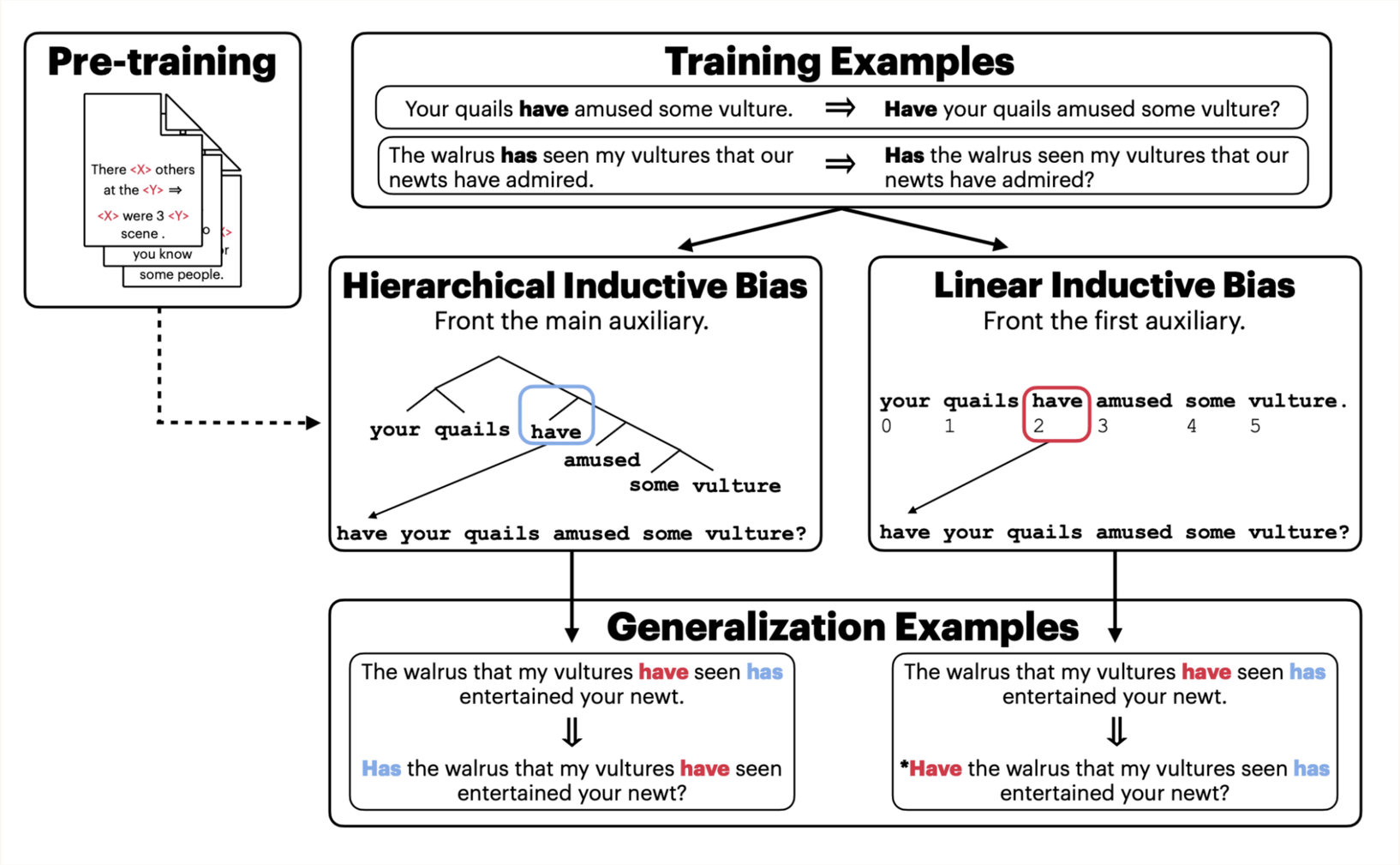
Lexical Generalization: use a known word in a new context

S.3.2. Novel Combination Modified Phrases and Grammatical Roles		
Object modification → Subject modification	Noah ate **the cake on the plate.**	**The cake on the table** burned.

Structural Generalization: extend known structures to be more complicated

S.3.3. Deeper Recursion		
Depth generalization: Sentential complements	Emma said **that** Noah knew **that** the cat danced.	Emma said **that** Noah knew **that** Lucas saw **that** the cat danced.
Depth generalization: PP modifiers	Ava saw the ball **in the bottle on the table**.	Ava saw the ball **in the bottle on the table on the floor**.

Appendix B: Passivization



Appendix C: BigBench Tasks

```
What is 42812 times 51570?

(a) 8127851667
(b) 9415523281
(c) 47549647
(d) 854486182733
(e) banana
(f) house
(g) 2207814840
```

```
What is the common morpheme among
these words: biology, antibiotic,
symbiosis, amphibian?

(a) disease
(b) life
(c) study
(d) animal
```

```
On which continent would
one find the Nile River?

(a) Africa
(b) Asia
(c) South America
(d) North America
(e) Europe
```

```
A man is offered one wish and he says
that he wants to be fire proof. His
wish is granted. The same man and a
woman are in a house. What will happen
if the house lights up on fire? Will
the man be hurt by the fire?

(a) Yes
(b) No
```

Appendix D: Permutation Tests

Could the BB results have arisen due to random chance?