How Do Transformers Learn Variable Binding in Symbolic Programs?

Yiwei Wu, Atticus Geiger, Raphaël Millière

Yiwei Wu

DLCT | August 15, 2025

Variable binding as a core computation

Back in the olden days

EITHER...

Connectionist models lack the kind of structured representations and structure-sensitive processes that can account for the systematicity of cognition

Fodor & Pylyshyn 1988

Back in the olden days

…OR

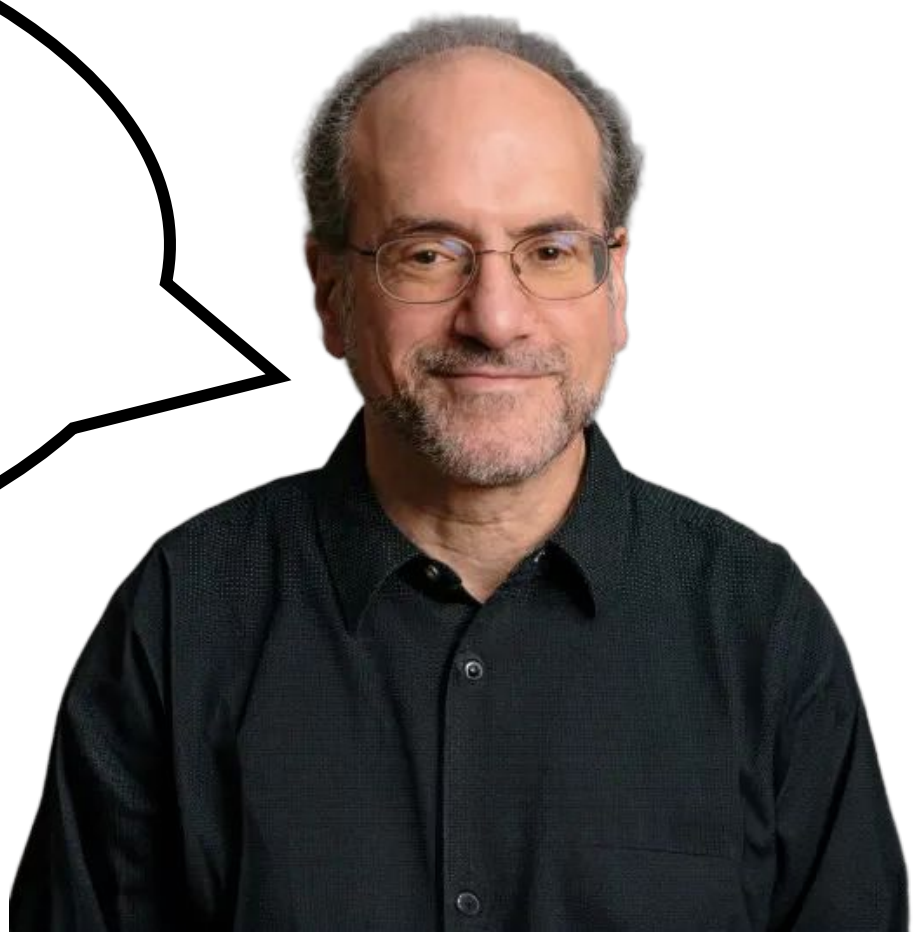
They do incorporate these but merely implement a classical symbol-processing architecture

Fodor & Pylyshyn 1988

Back in the olden days

NEITHER/NOR!

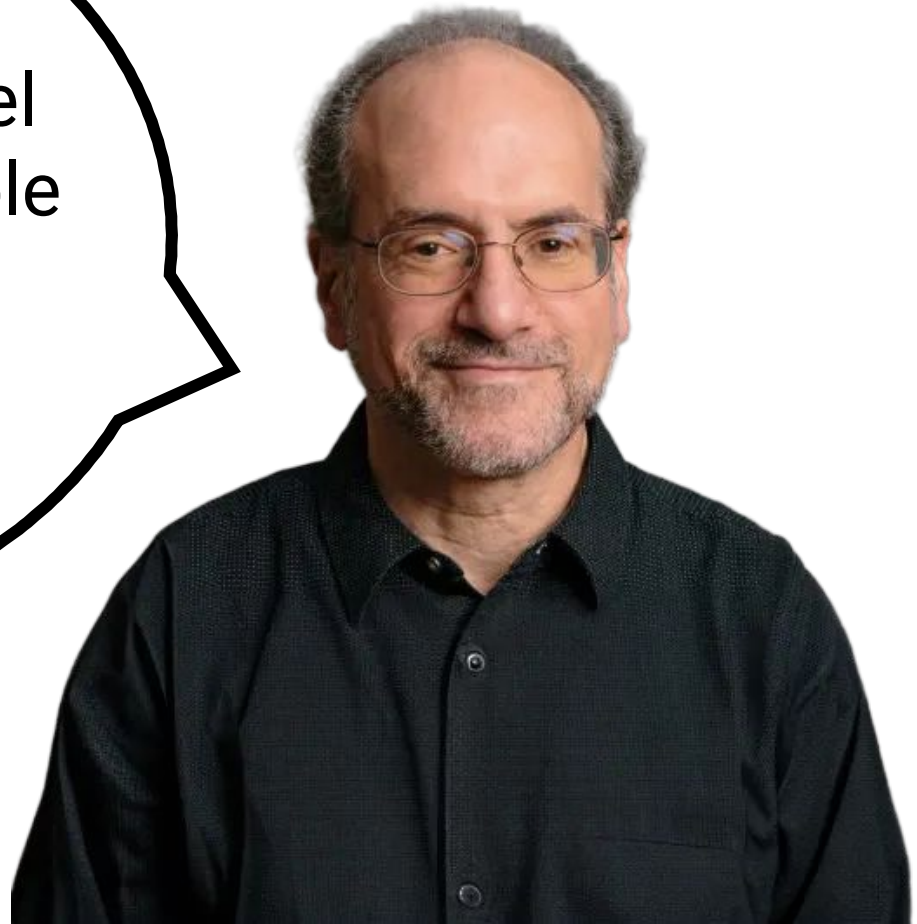
Neural networks can have the requisite structure without implementing a classical architecture



Smolensky 1988

Back in the olden days

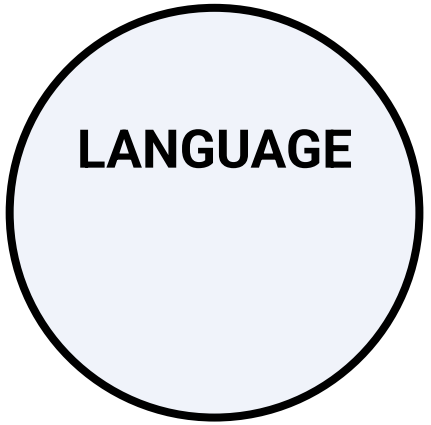
Any adequate model must provide a viable mechanism for variable binding.



Variable binding

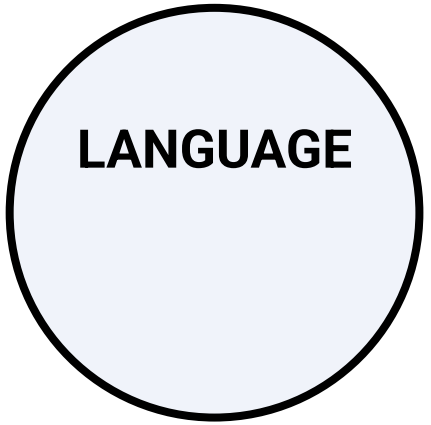
The process of associating a variable (placeholder, role) with a specific value (instance, filler) within a structured representation, such that the value can be dynamically updated and retrieved for use in downstream computations.

Variable binding



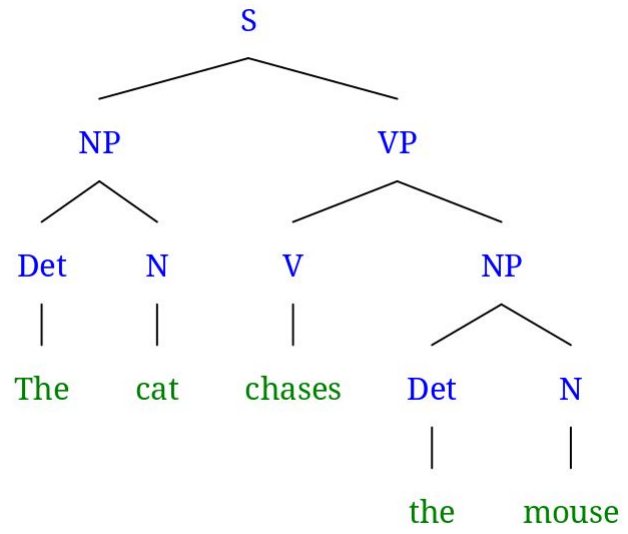
Anaphora	John$_i$ saw his$_i$ dog.
Quantification	Every student$_i$ read a book that they$_i$ liked.
Wh-Movement	Who$_i$ did Mary see $\_\_{}_i$?

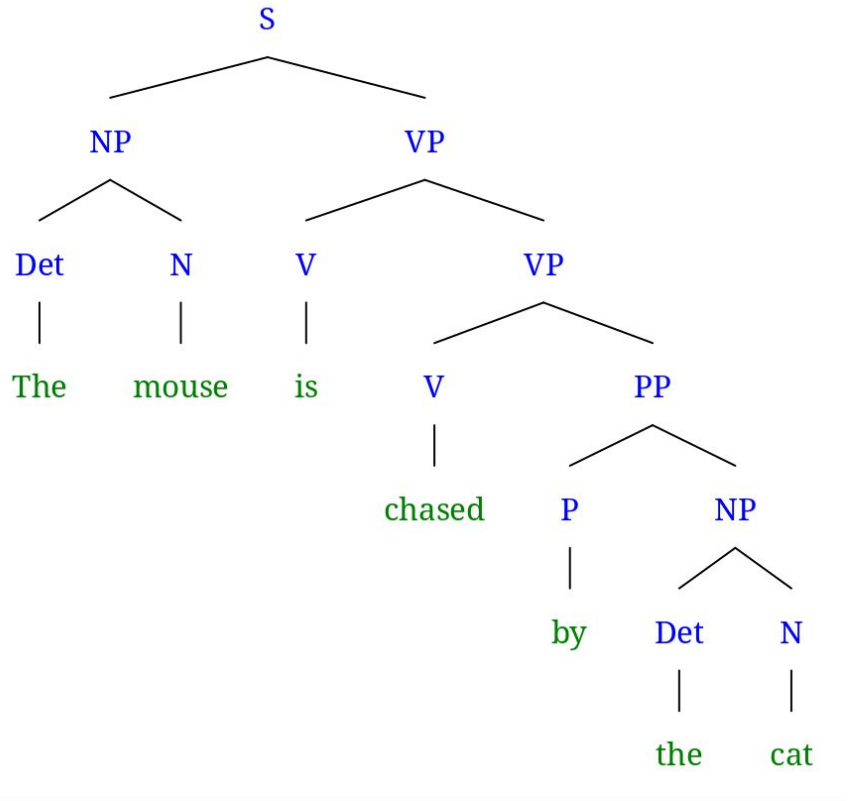
Variable binding



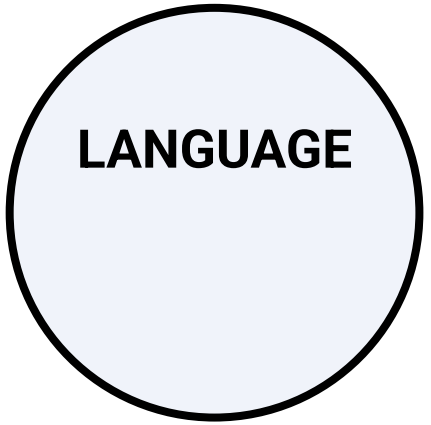
The cat chases the mouse.

The mouse is chased by the cat.





Variable binding



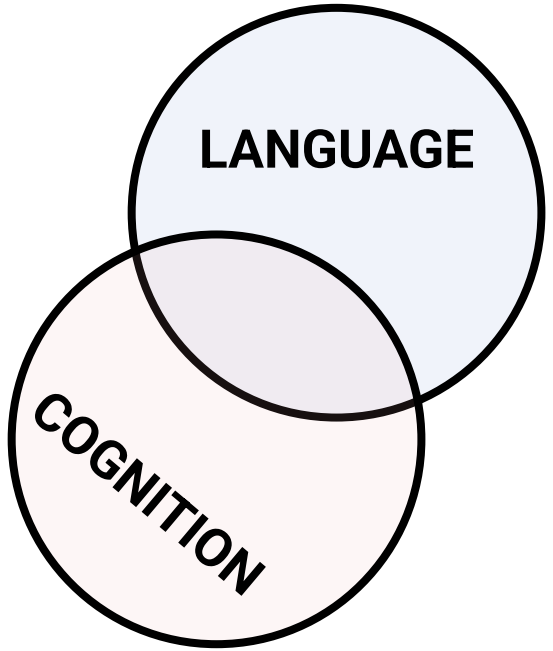
The cat chases the mouse.

The mouse is chased by the cat.

Logical form $\exists x \exists y[\text{CAT}(x) \wedge \text{MOUSE}(y) \wedge \text{CHASE}(x,y)]$

Thematic roles $\text{AGENT}(\text{CAT}), \text{THEME}(\text{MOUSE})$

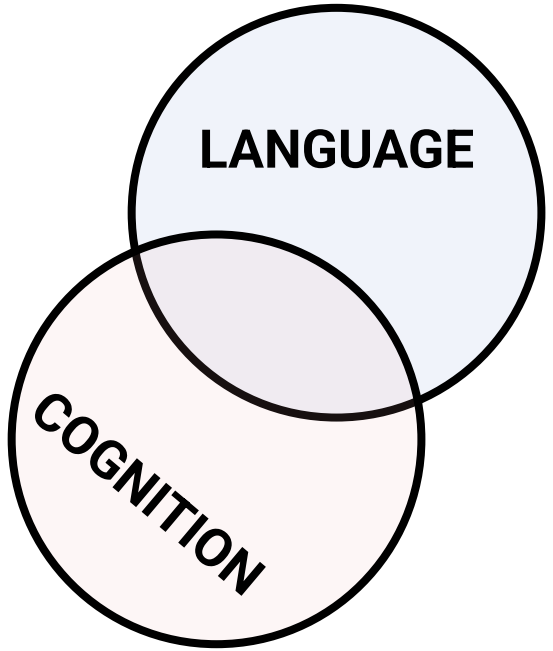
Variable binding



Marcus 2001; Gallistel & King 2011

Variable binding

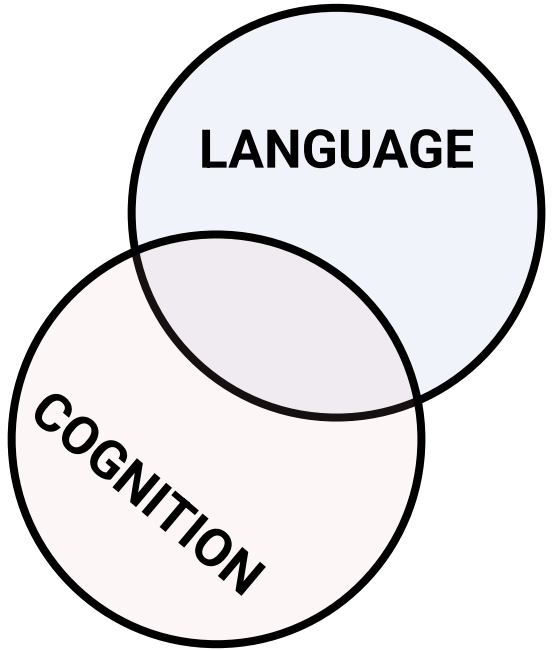
- Systematicity & compositional generalization



Marcus 2001; Gallistel & King 2011

Variable binding

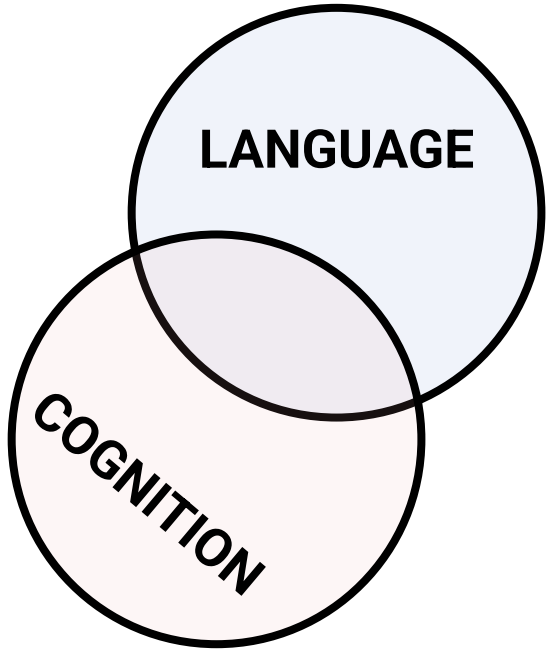
- Systematicity & compositional generalization
- Rule-based learning



Marcus 2001; Gallistel & King 2011

Variable binding

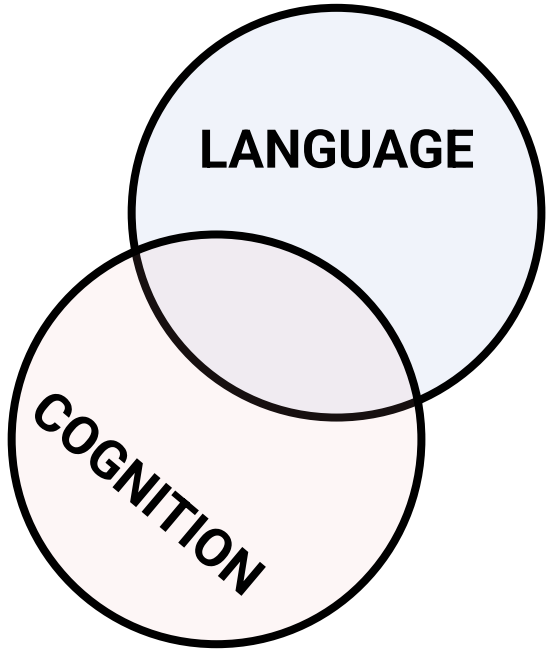
- Systematicity & compositional generalization
- Rule-based learning
- Abstract role-based reasoning



Marcus 2001; Gallistel & King 2011

Variable binding

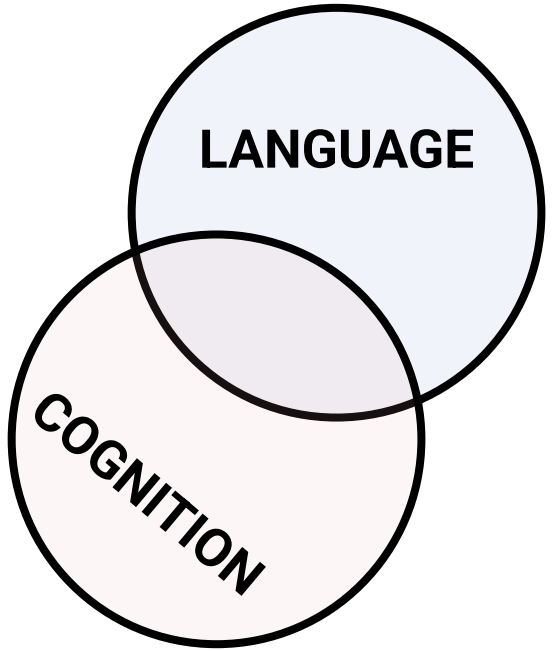
- Systematicity & compositional generalization
- Rule-based learning
- Abstract role-based reasoning
- Analogical reasoning



Marcus 2001; Gallistel & King 2011

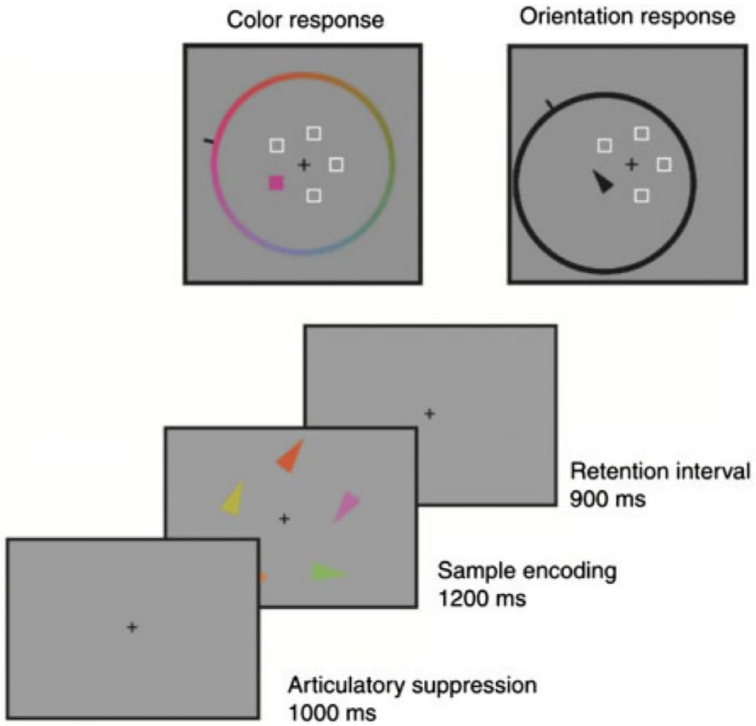
Variable binding

- Systematicity & compositional generalization
- Rule-based learning
- Abstract role-based reasoning
- Analogical reasoning
- Event understanding



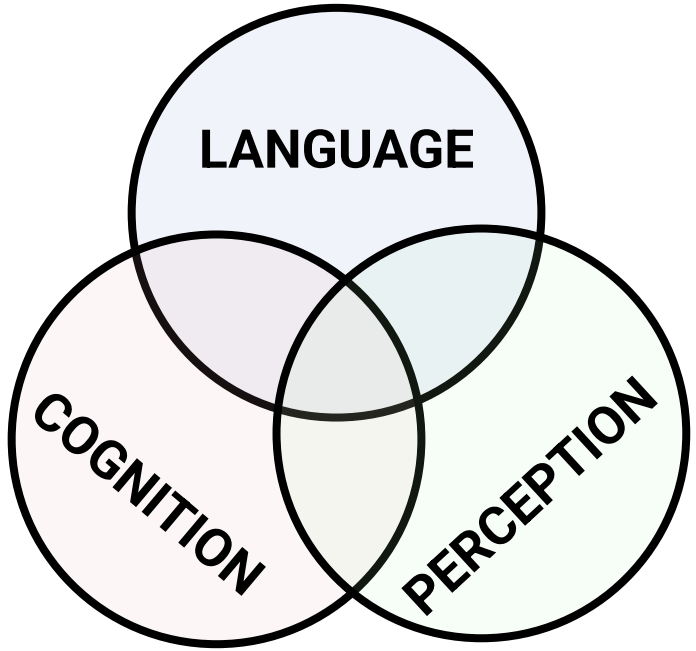
Marcus 2001; Gallistel & King 2011

Variable binding



Object files

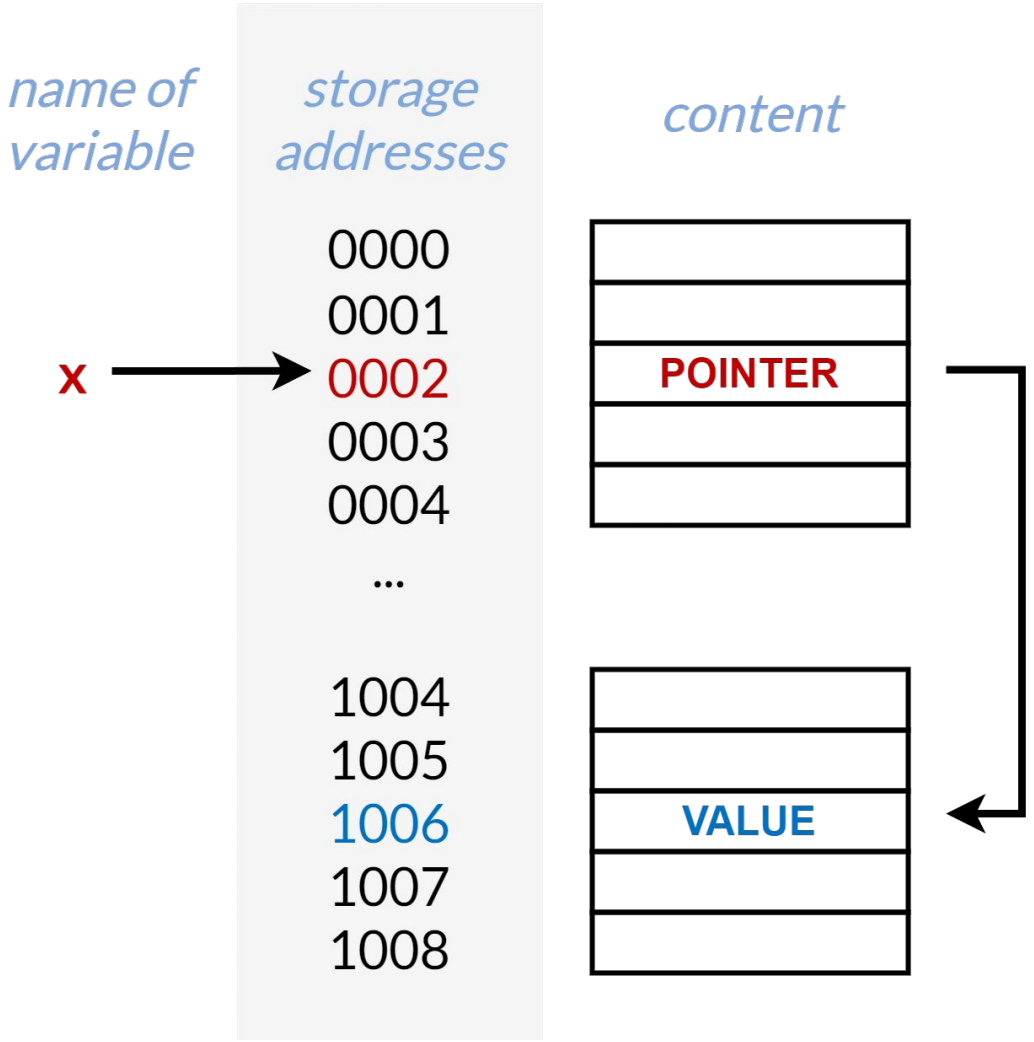
Abstract relations



Green & Quilty-Dunn 2021; Hafri et al. 2024

Indirect addressing

- Variable binding is classically implemented through indirect addressing
- The first address serves as a symbol for the variable, pointing to the location containing the address of the value
- The actual value is specified by the bit pattern at the second address, which is indirectly accessed



Marcus (2001), Gallistel and King (2010), Bottou (2014)

Modern DNNs

"Variable binding [is] a classic example of LoT-like symbolic computation"

"It remains open that DNNs might mimic the performance of biological perception and cognition across a wide variety of domains and tasks by *implementing* core features of LoTs."

Quilty-Dunn et al. 2023

Two questions

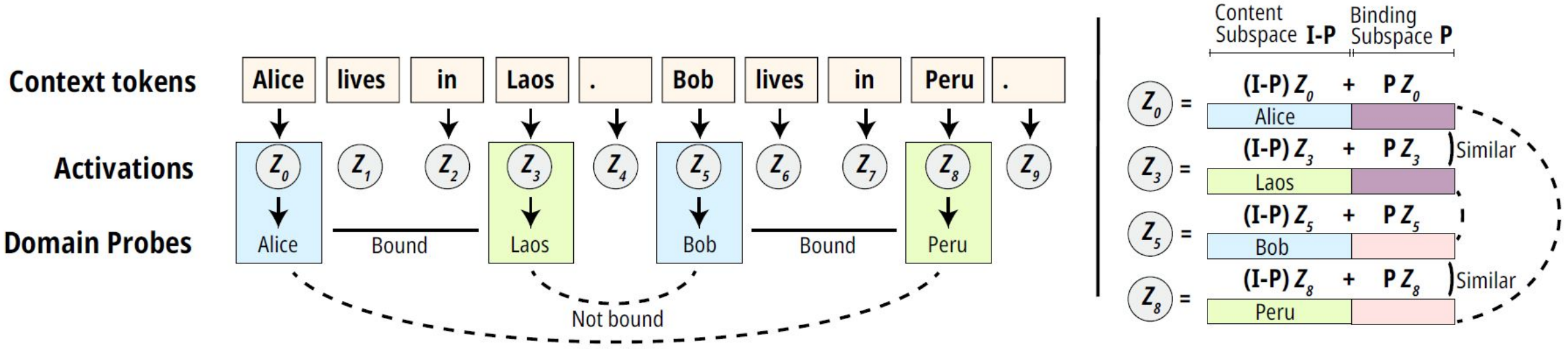
Can Transformers behave consistently with the hypothesis that they have a mechanism for variable binding?

If so, how does this mechanism work, and how does it emerge?

2

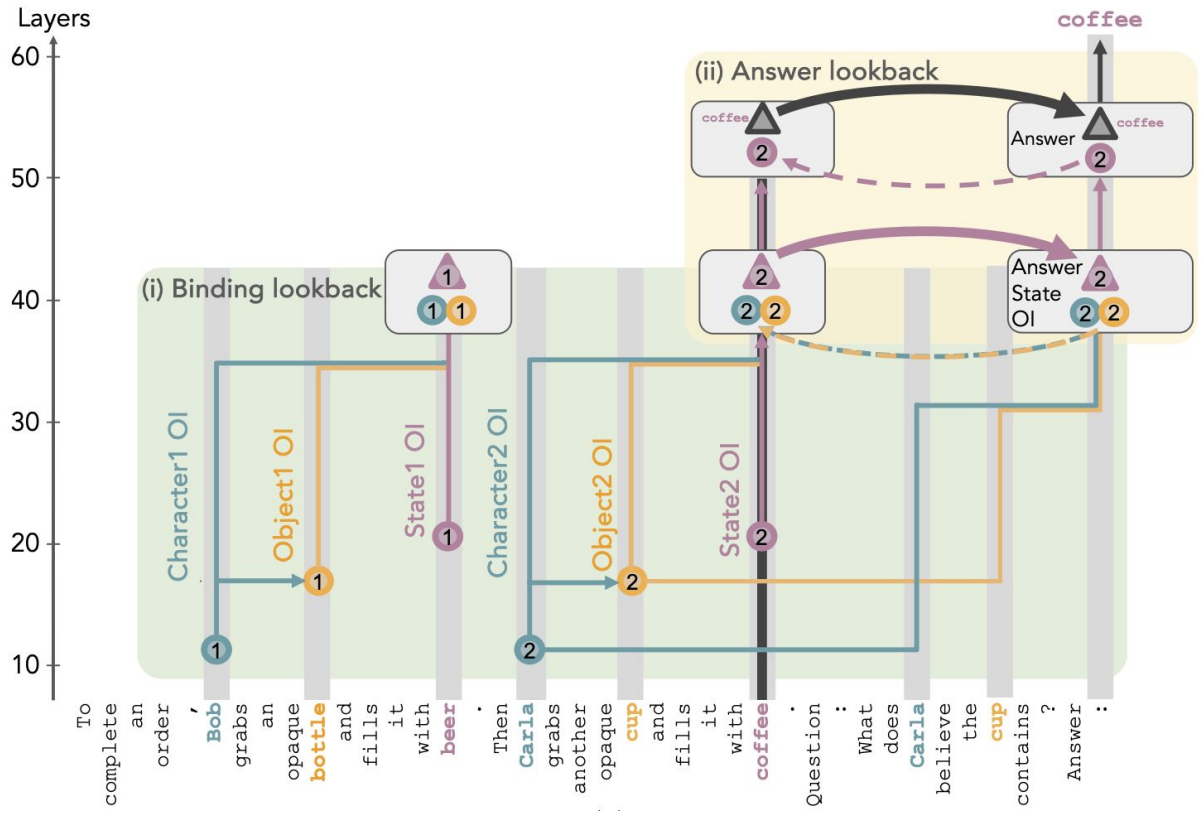
A developmental & mechanistic perspective

Related work: entity binding in pretrained LLMs



Feng & Steinhardt 2024; Feng et al. 2024; Dai et al. 2024

Related work: entity binding in pretrained LLMs



Counterfactual

Carla and Bob are working in a busy restaurant. To complete an order, Carla grabs an opaque cup and fills it with tea. Then Bob grabs another opaque bottle and fills it with water. Question: What does Carla believe the cup contains? Answer: tea

Original

Bob and Carla are working in a busy restaurant. To complete an order, Bob grabs an opaque bottle and fills it with beer. Then Carla grabs another opaque cup and fills it with coffee. Question: What does Carla believe the cup contains? Answer: coffee

Intervention 1: Answer Pointer (●), ***Causal Model Output:*** beer
Intervention 2: Answer Payload (▲), ***Causal Model Output:*** tea

Prakash et al. 2025

Atticus Geiger

Raphaël Millière

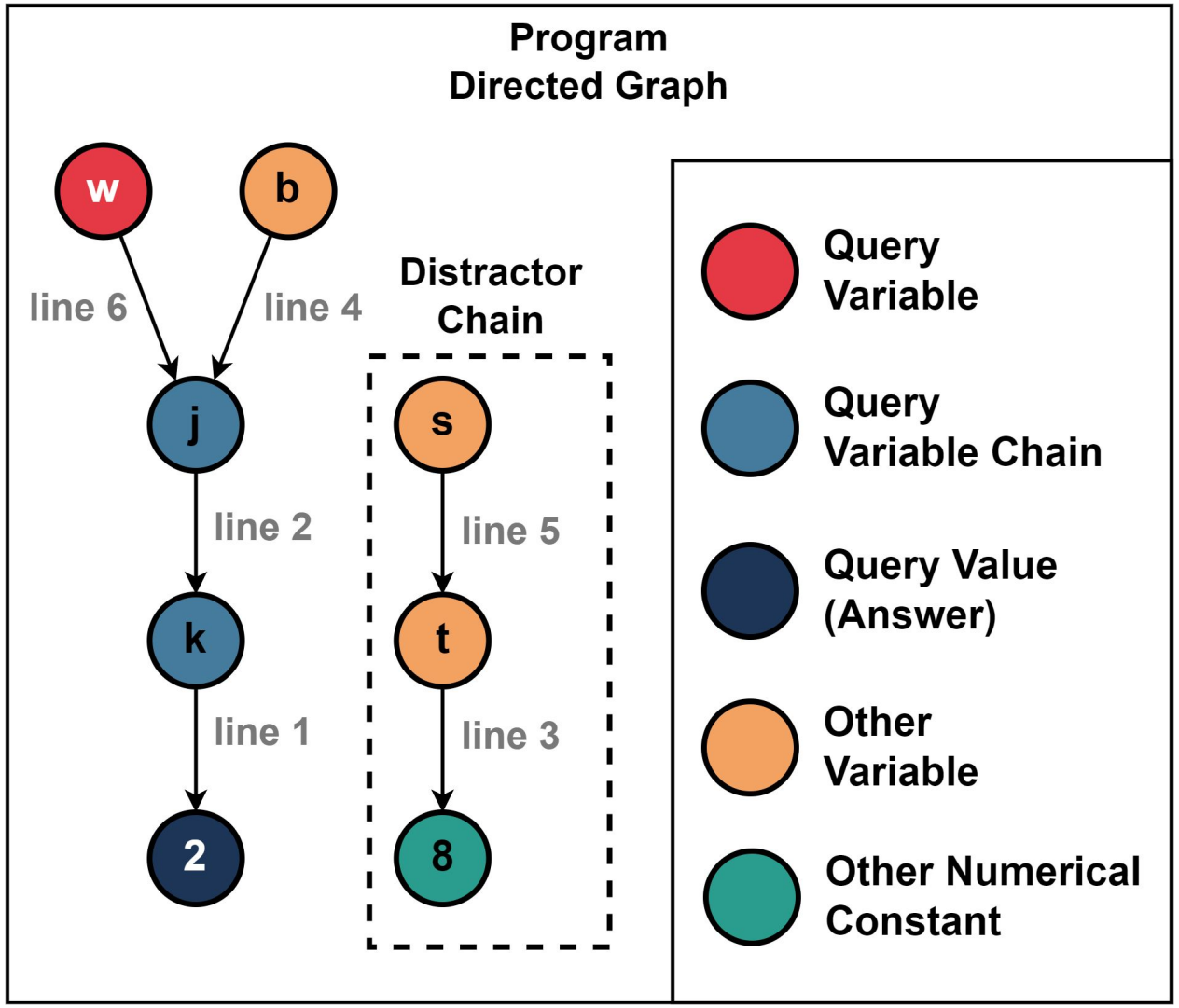
The experiment

- **Setup**: we train a small Transformer-based language model on a synthetic variable binding task with causal language modeling objective
- **Behavioral component**: we assess how performance on a held-out test set evolves over the course of training
- **Interpretability component**: we use probing and interventions to understand what strategy the model learns and how it learns it

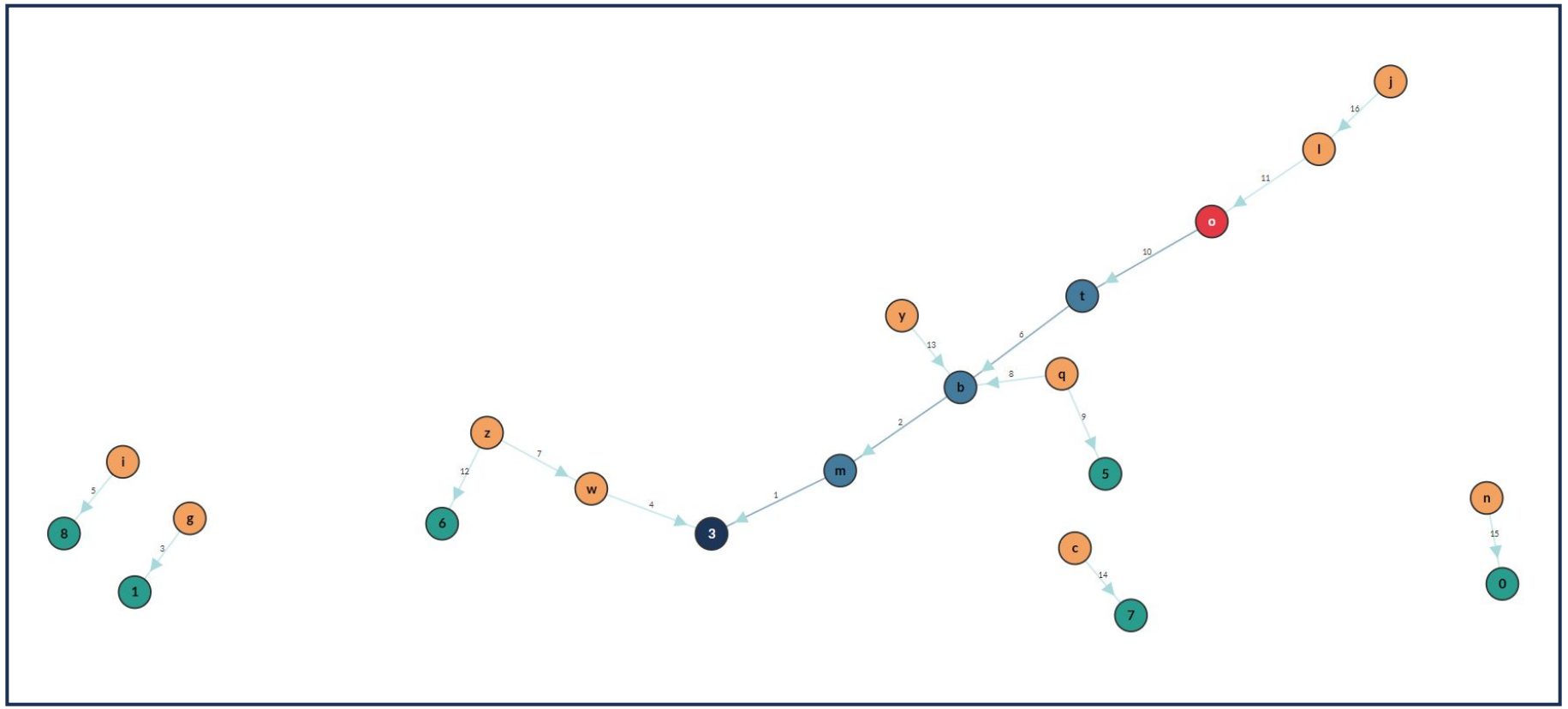
The task (abbr.)

Example 3-Hop Program

referential depth 1	`k=2`
referential depth 2	`j=k`
distractor chain	`t=8`
	`b=j`
	`s=t`
referential depth 3	`w=j`
query	`#w:`



The task (for real)



Actual 4-Hop Program

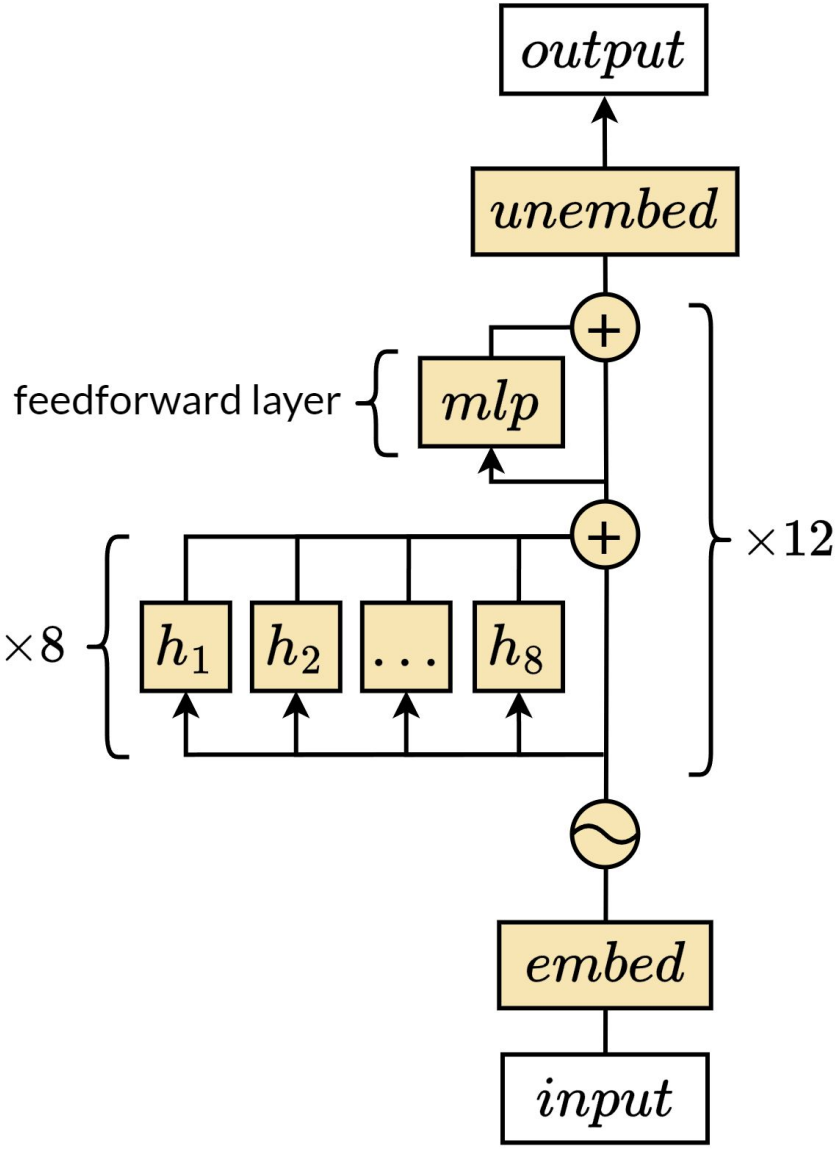
```
referential depth 1  m=3
referential depth 2  b=m
                     g=1
                     w=3
                     i=8
referential depth 3  t=b
                     z=w
                     q=b
                     q=5
referential depth 4  o=t
                     l=o
                     z=6
                     y=b
                     c=7
                     n=0
                     j=l
query                #o:
```

Sampling

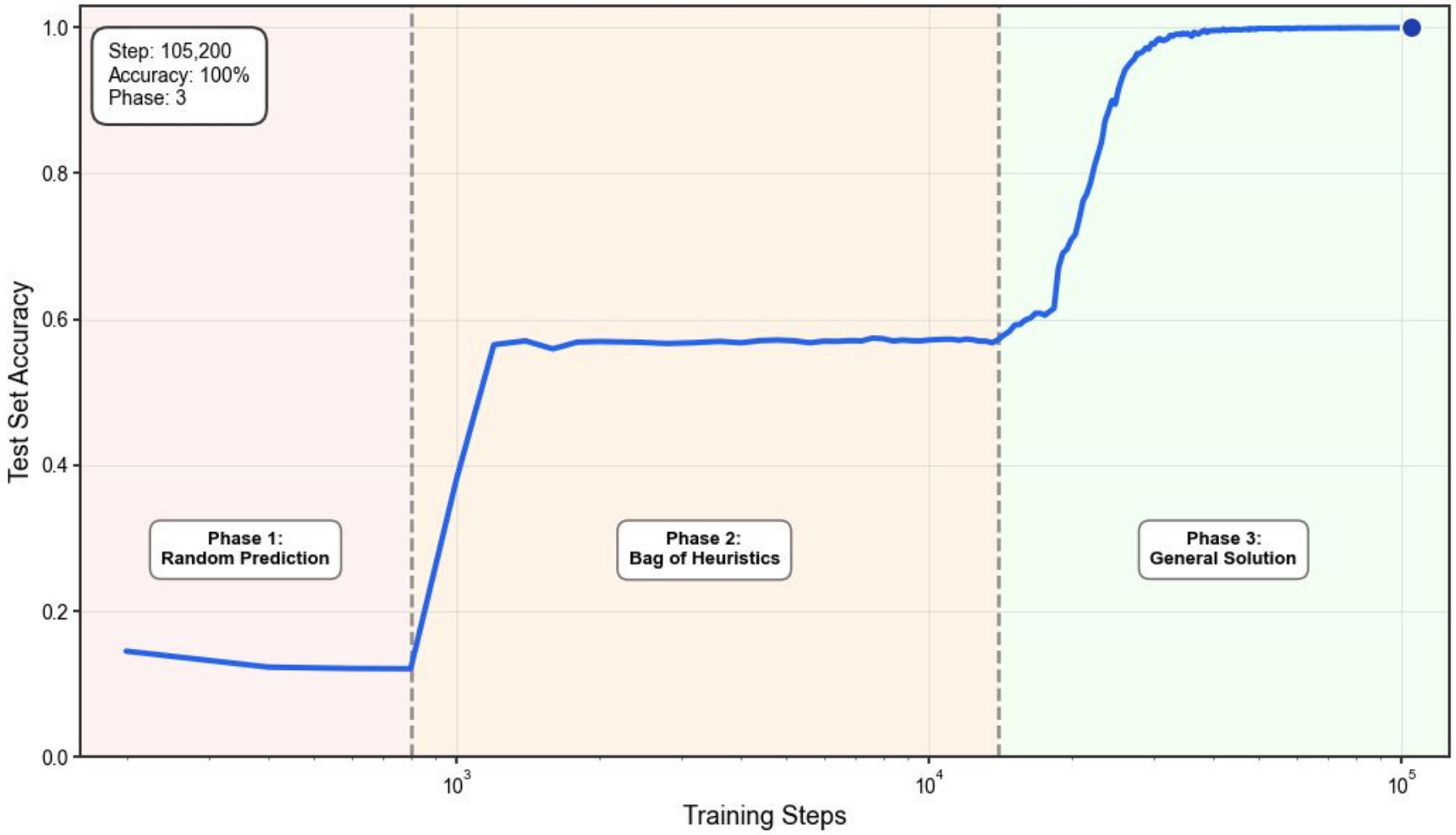
- 500,000 programs
- Data split: 90% train / 0.2% val / 9.8% test
- 26 variable names (a-z)
- 10 constants (0-9)
- We favor longer chains
- We use rejection sampling to balance the data across the 4 possible referential depths for the query variable chain

Model

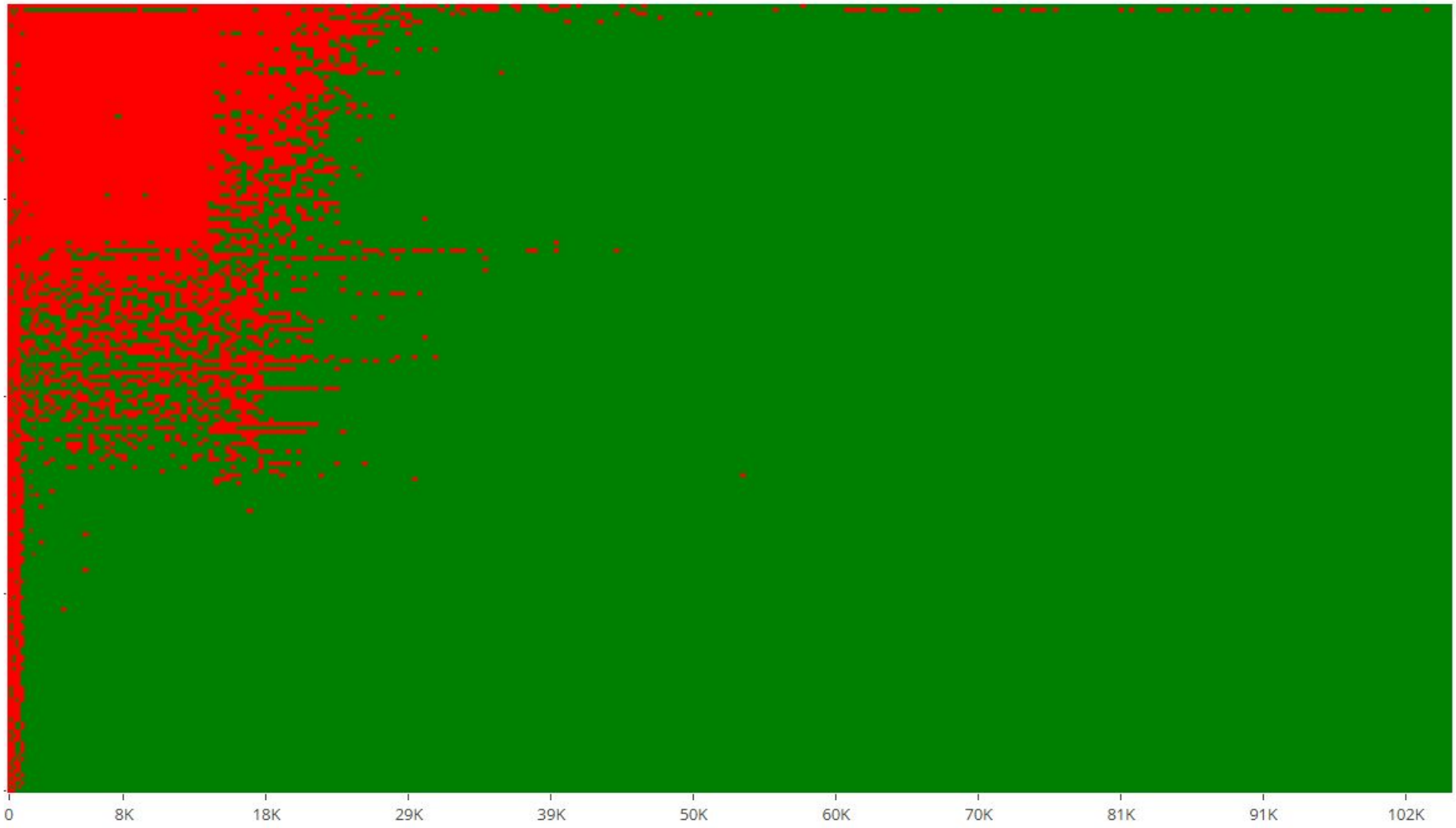
- Transformer architecture (GPT-2-like)
- 37.8M parameters
- 12 layers (embedding dim 512)
- 8 attention heads per layer (dim 64)
- Rotary positional embedding (RoPE)
- GELU activations
- Dropout rate: 0.1



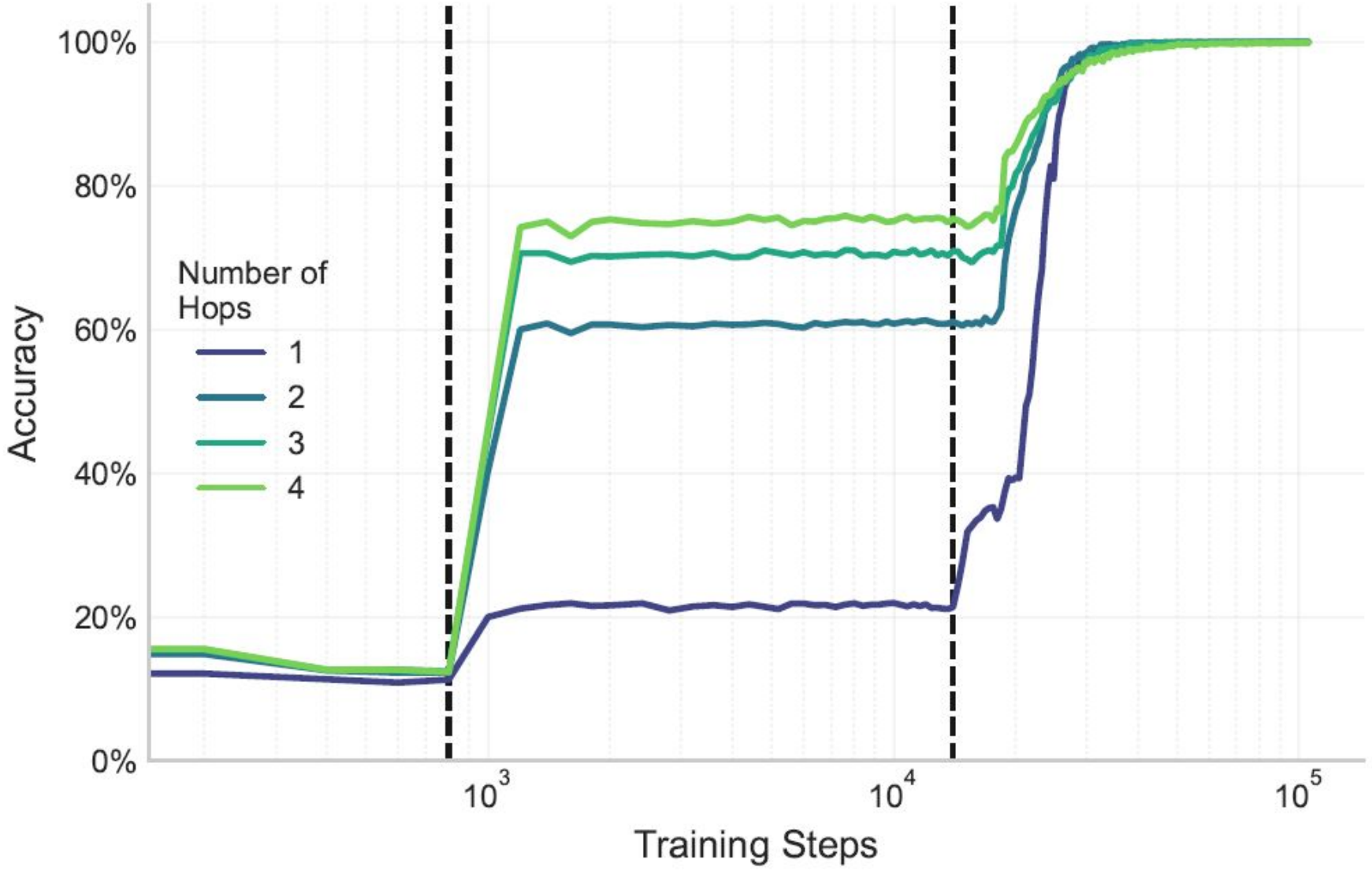
Behavioral evaluation



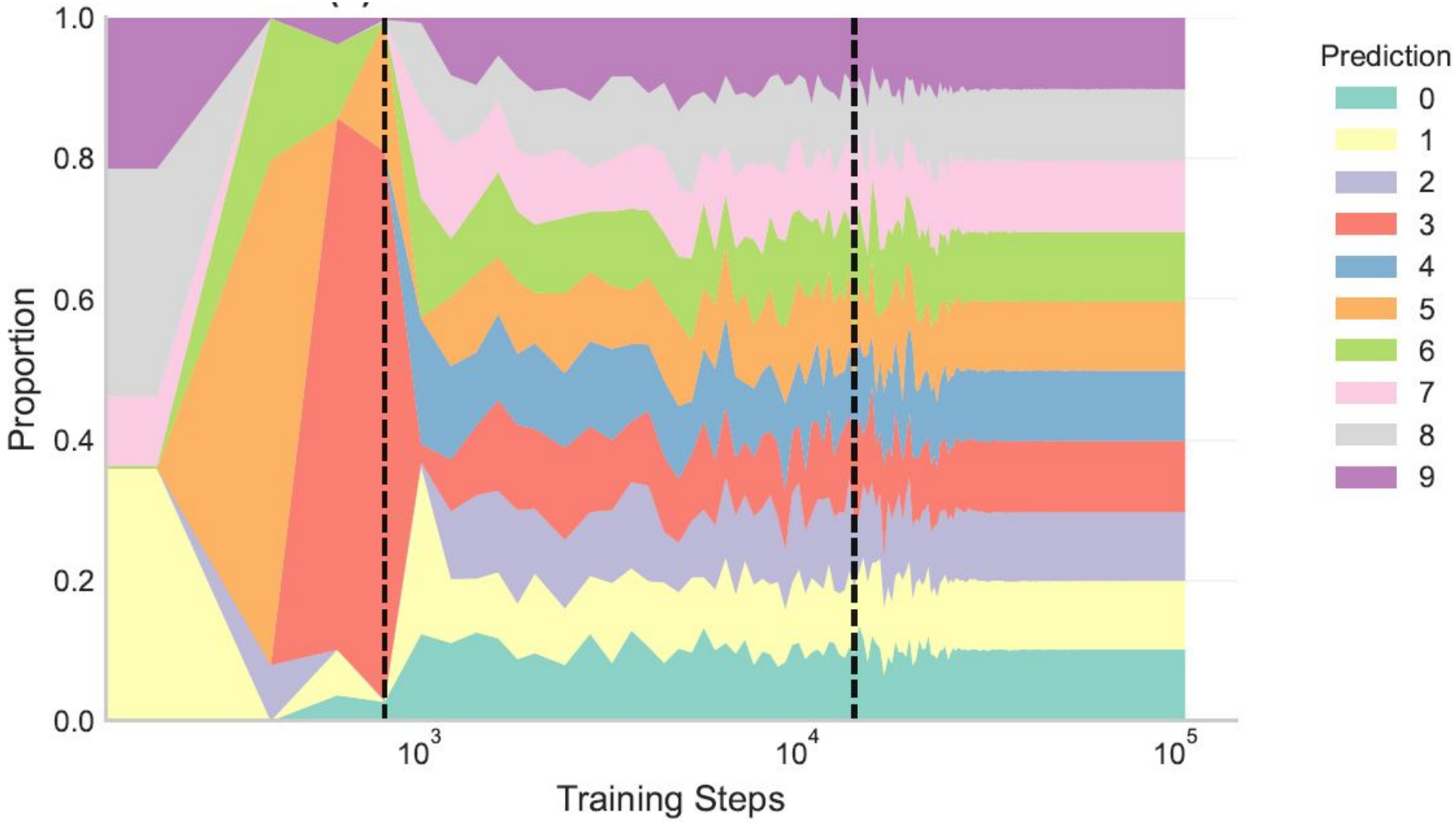
Behavioral evaluation



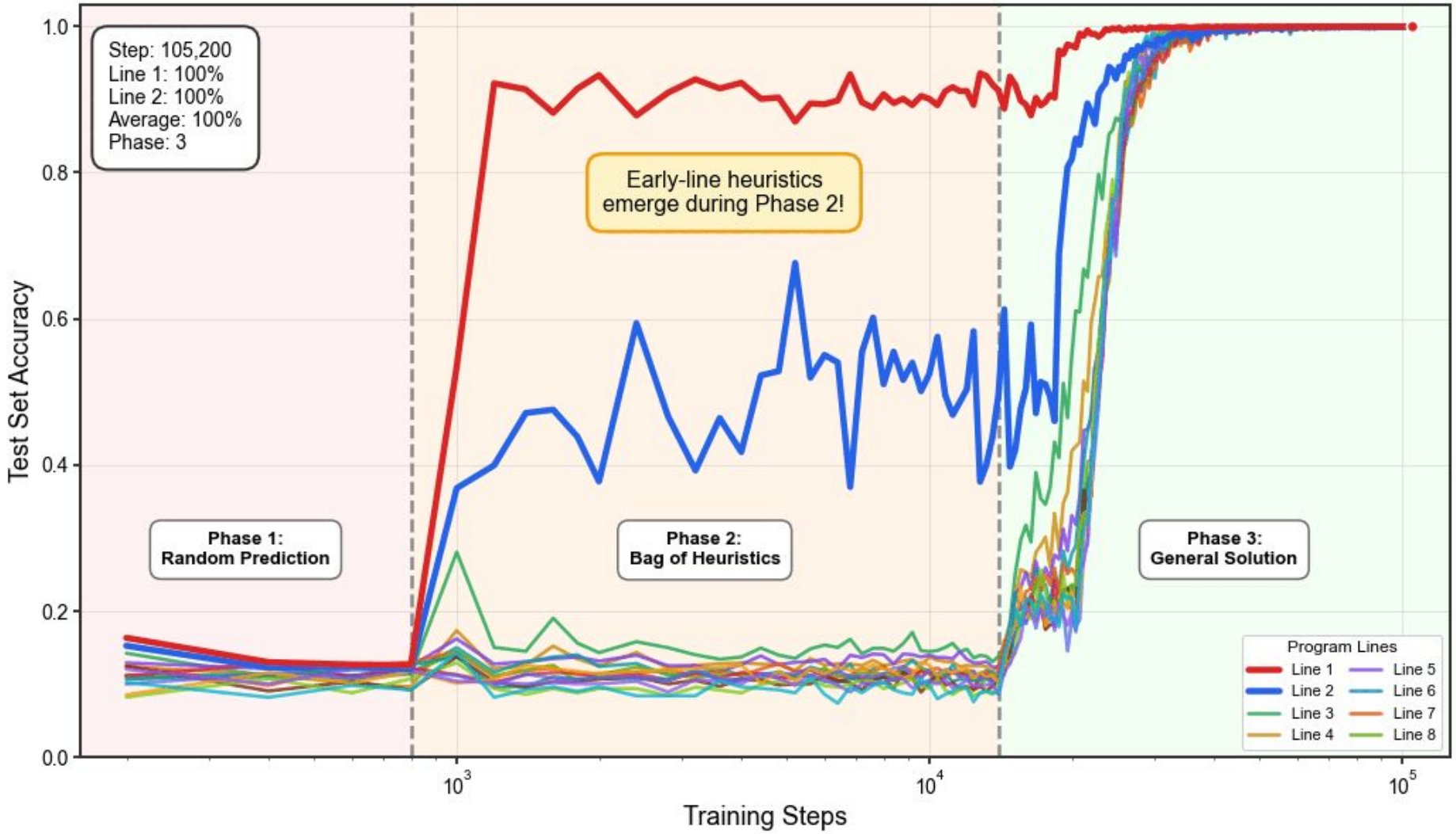
Behavioral evaluation



Behavioral evaluation

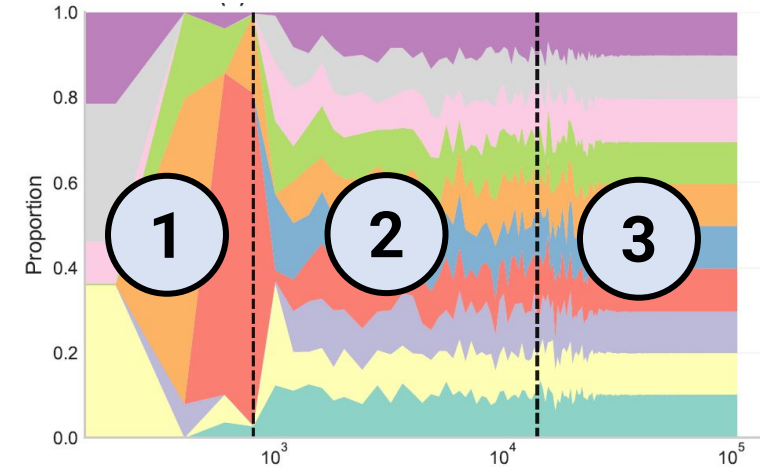


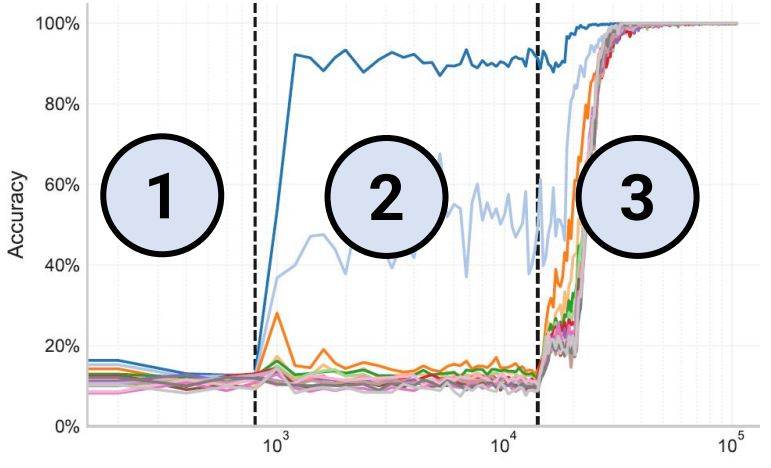
Behavioral evaluation

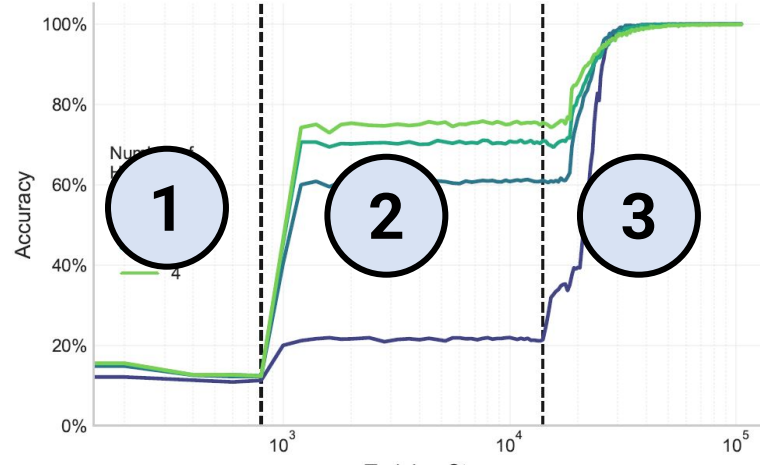


Behavioral evaluation

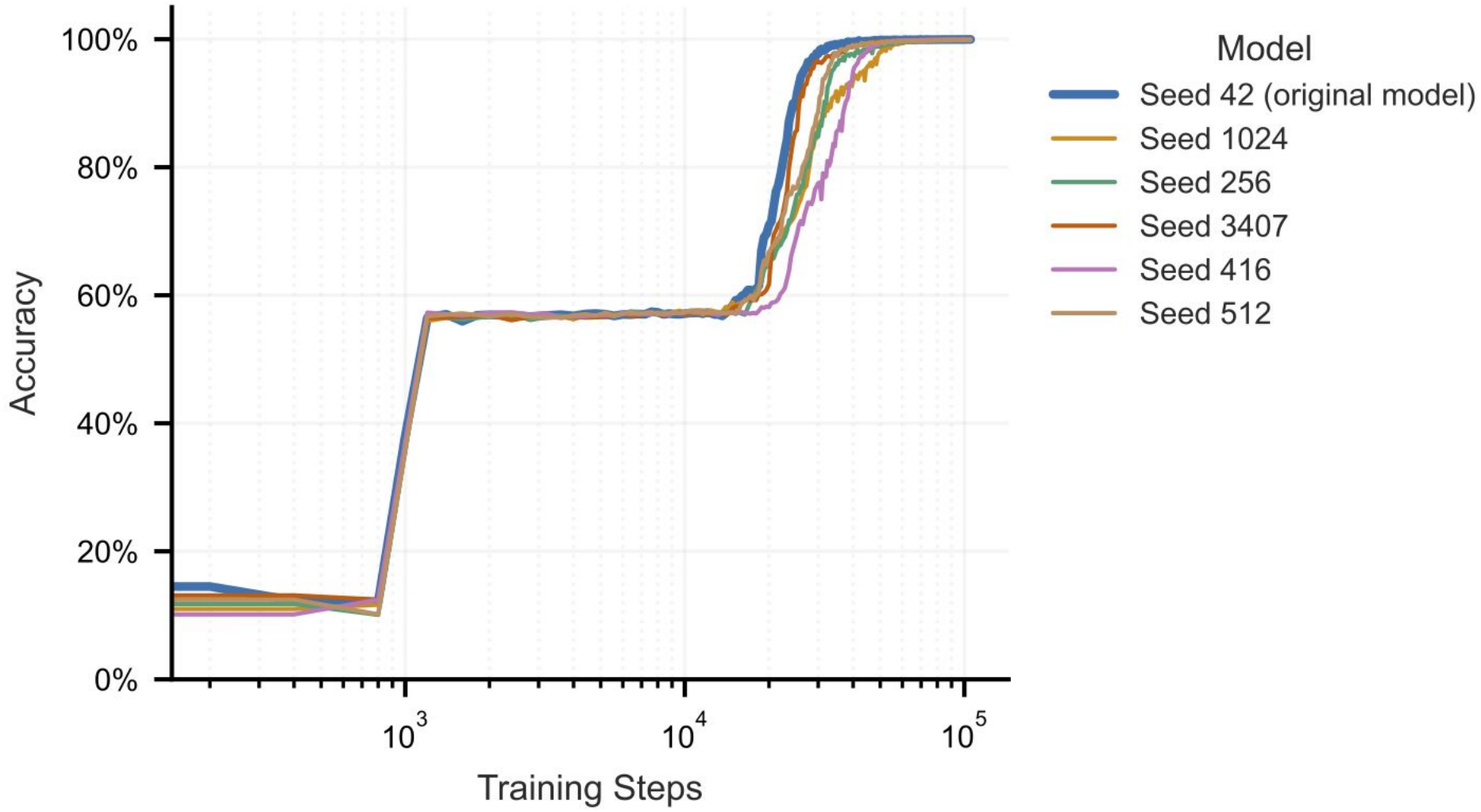
- **Phase 1**: predicting random constants
- **Phase 2**: bag of early-line heuristics
- **Phase 3**: systematic solution



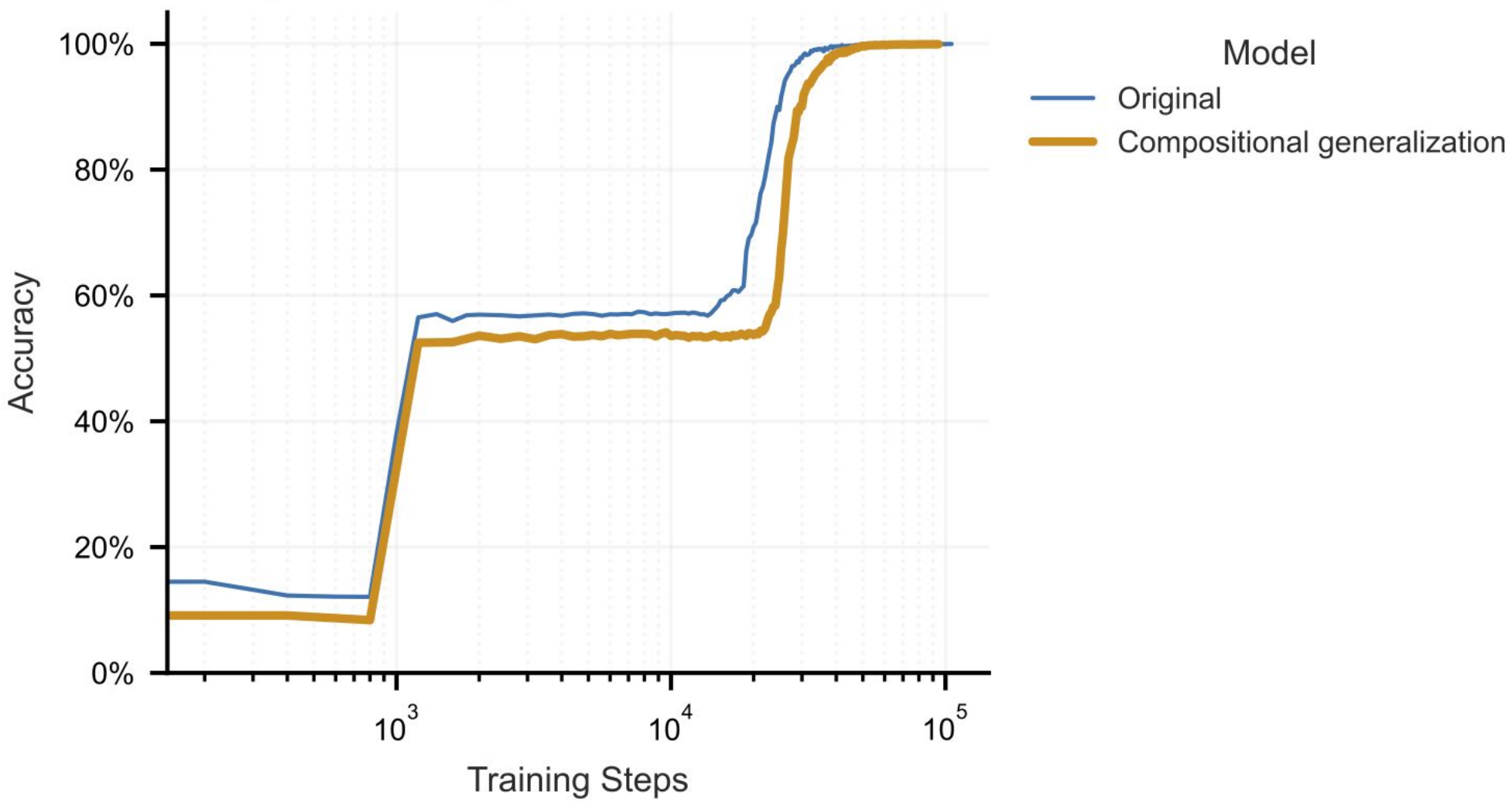




Multiple Random Seeds



Generalization to Unseen Combinations

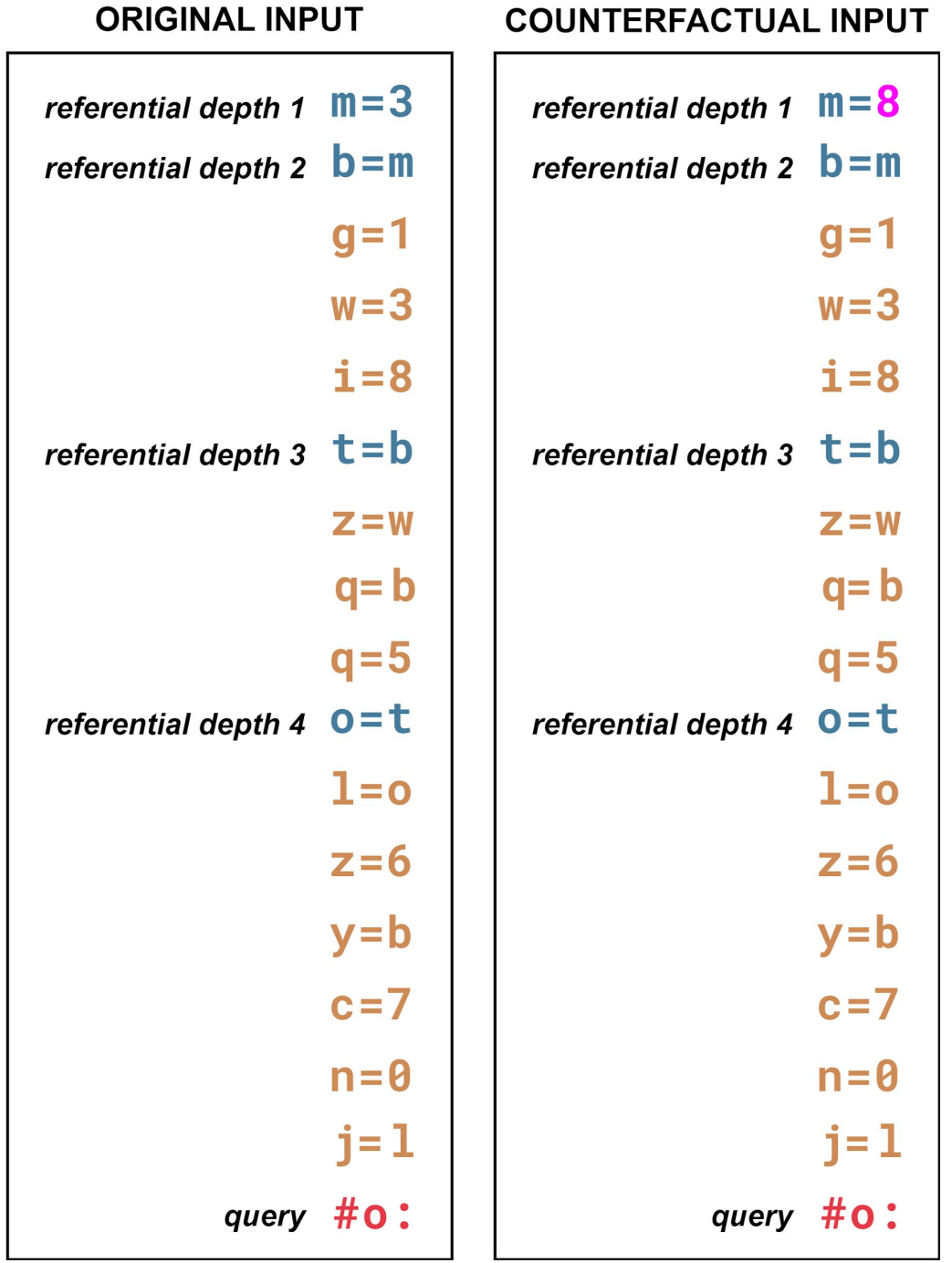


Probing experiment

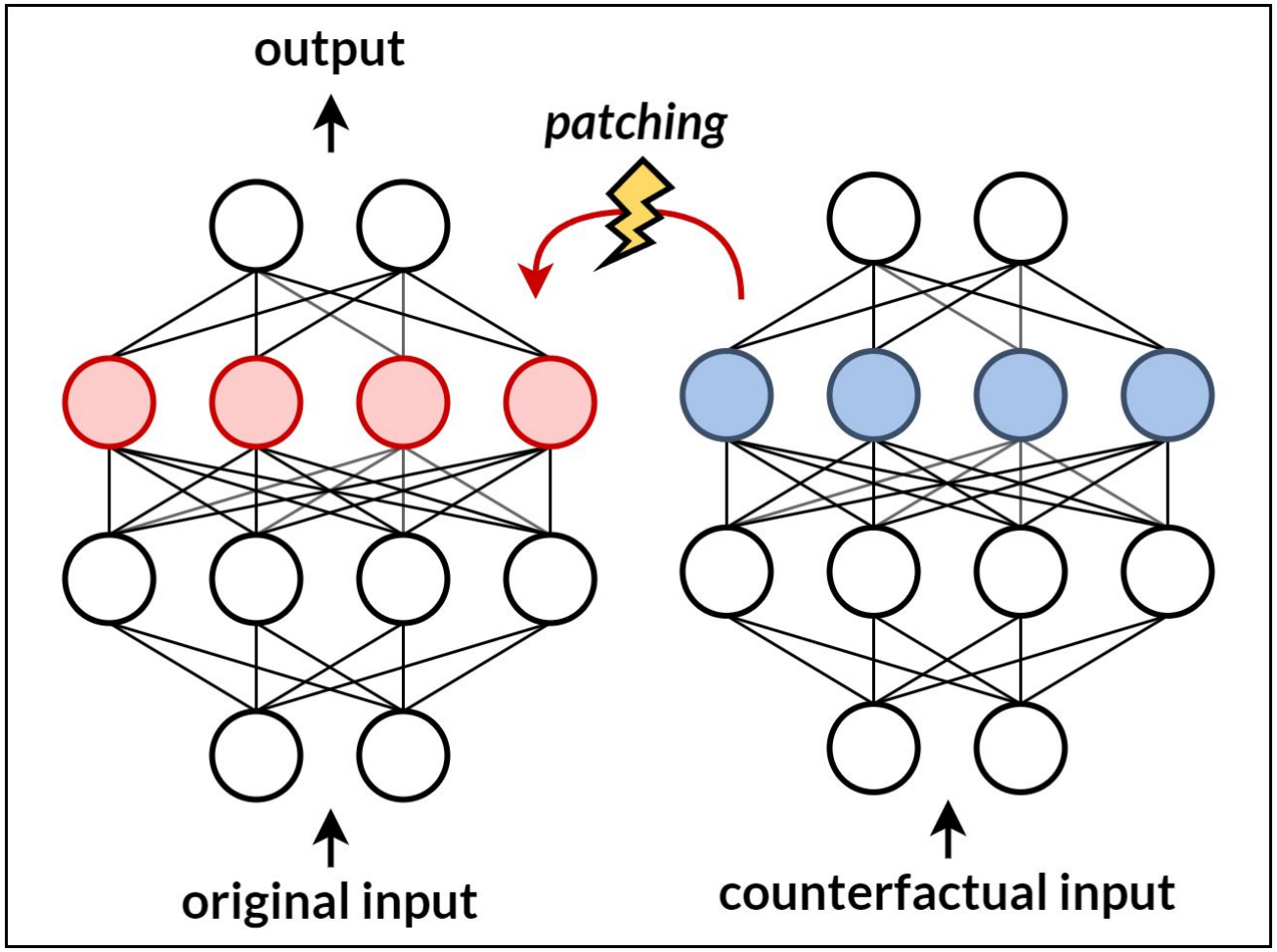
Layer	State Acc (%)	Var. Acc (Excl. Nil) (%)
1	7.71	21.28
2	8.42	25.36
3	8.78	28.56
4	8.87	28.52
5	8.73	29.80
6	**8.90**	**30.87**
7	8.88	30.72
8	8.83	30.03
9	8.85	29.61
10	8.73	28.90
11	8.72	28.66
12	8.77	28.51

Interchange interventions

- Sample a program (original input)
- Create a counterfactual input with a different root value for query chain
- Cache model activations on counterfactual input
- Swap activations of specific model components on original input with cached activations
- Track effect on logits and behavior

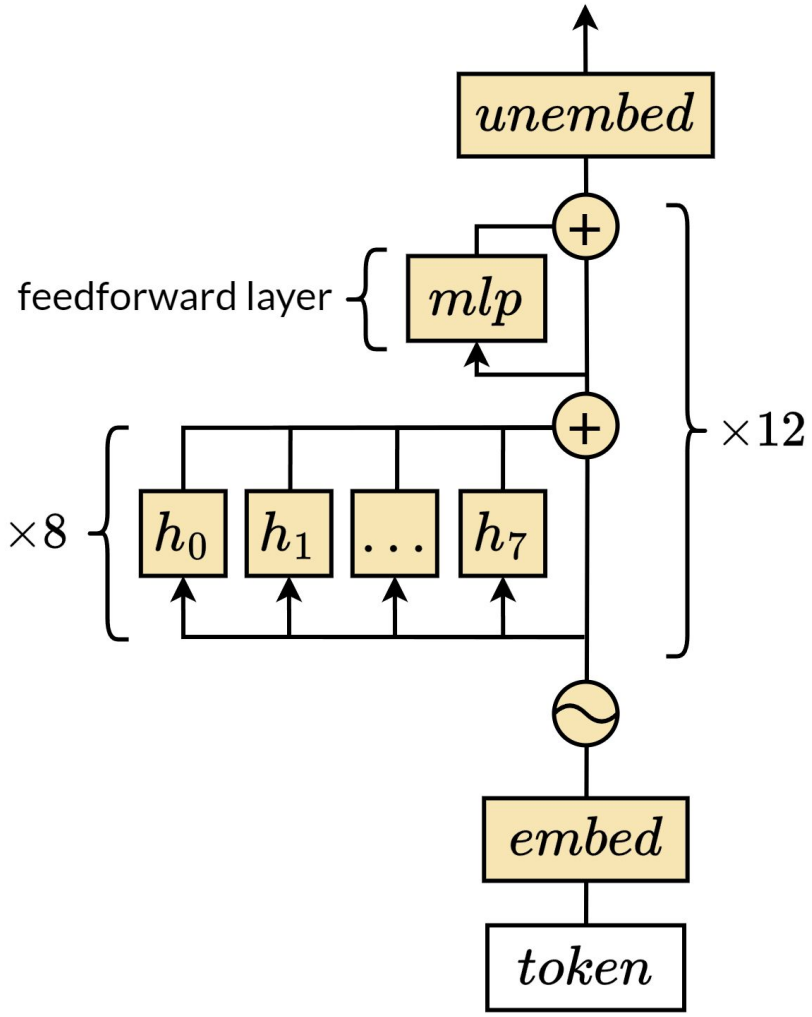


Interchange interventions

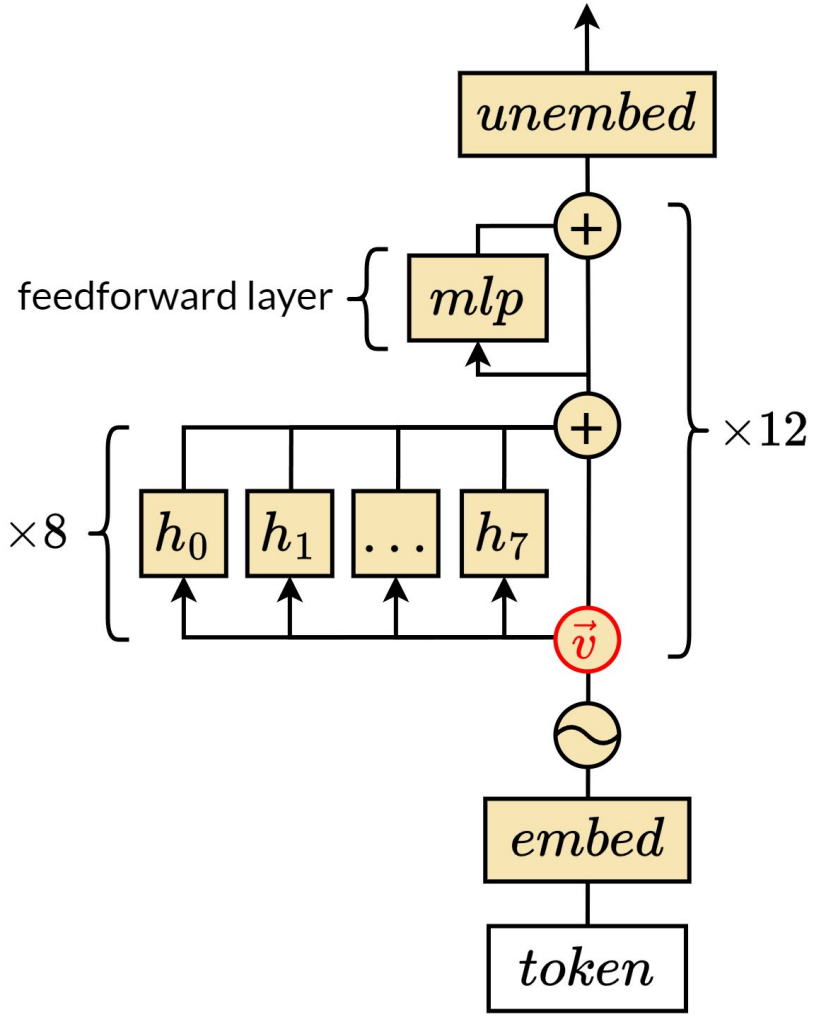


	ORIGINAL INPUT	COUNTERFACTUAL INPUT
referential depth 1	m=3	m=8
referential depth 2	b=m	b=m
	g=1	g=1
	w=3	w=3
	i=8	i=8
referential depth 3	t=b	t=b
	z=w	z=w
	q=b	q=b
	q=5	q=5
referential depth 4	o=t	o=t
	l=o	l=o
	z=6	z=6
	y=b	y=b
	c=7	c=7
	n=0	n=0
	j=l	j=l
query	#o:	#o:

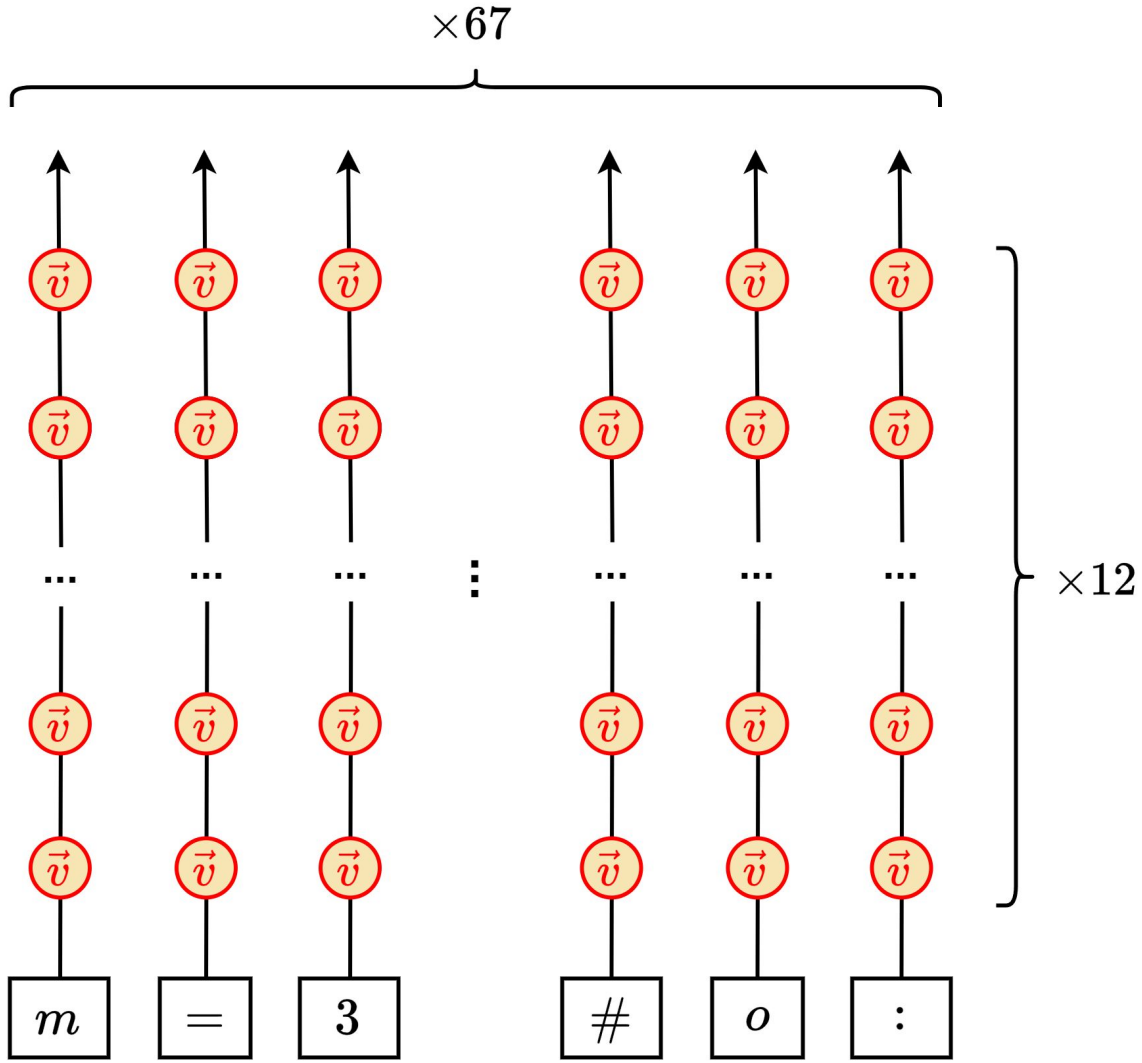
Patching the residual stream per layer

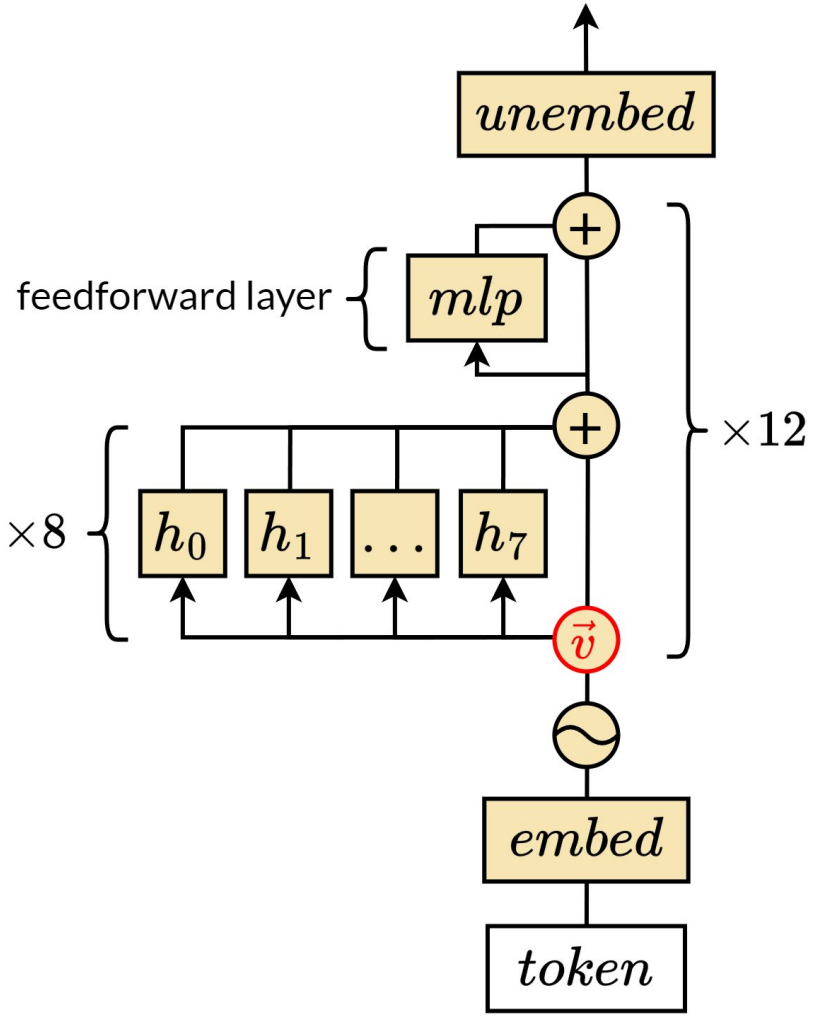


Patching the residual stream per layer

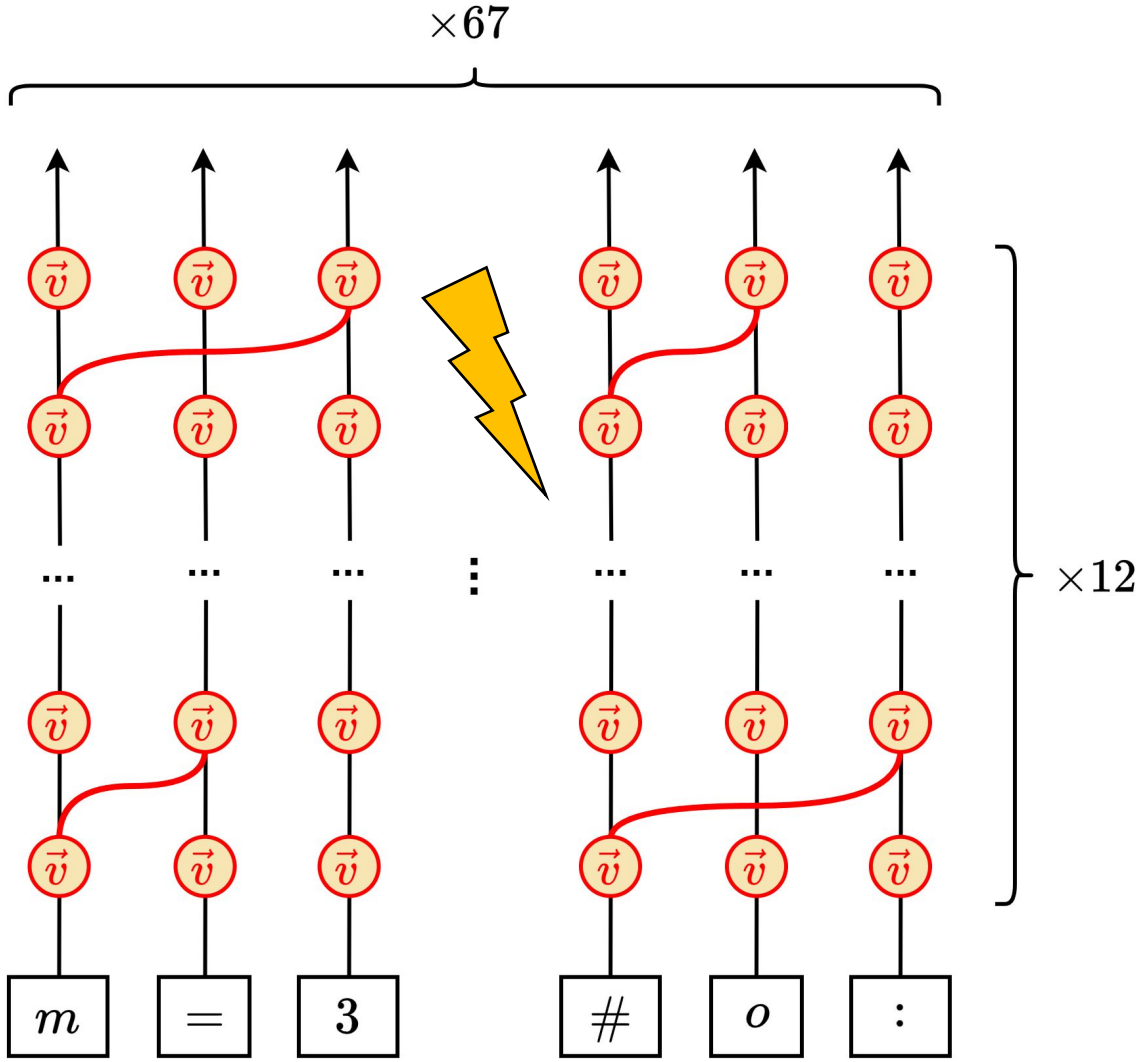


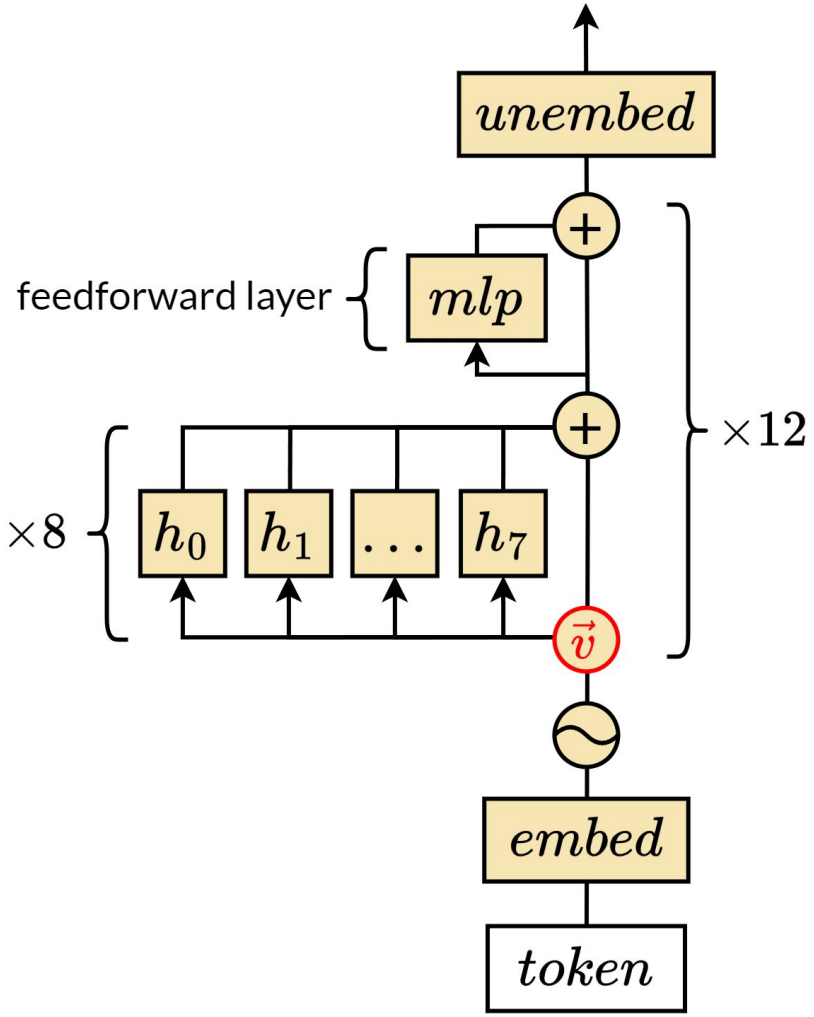
Patching the residual stream per layer



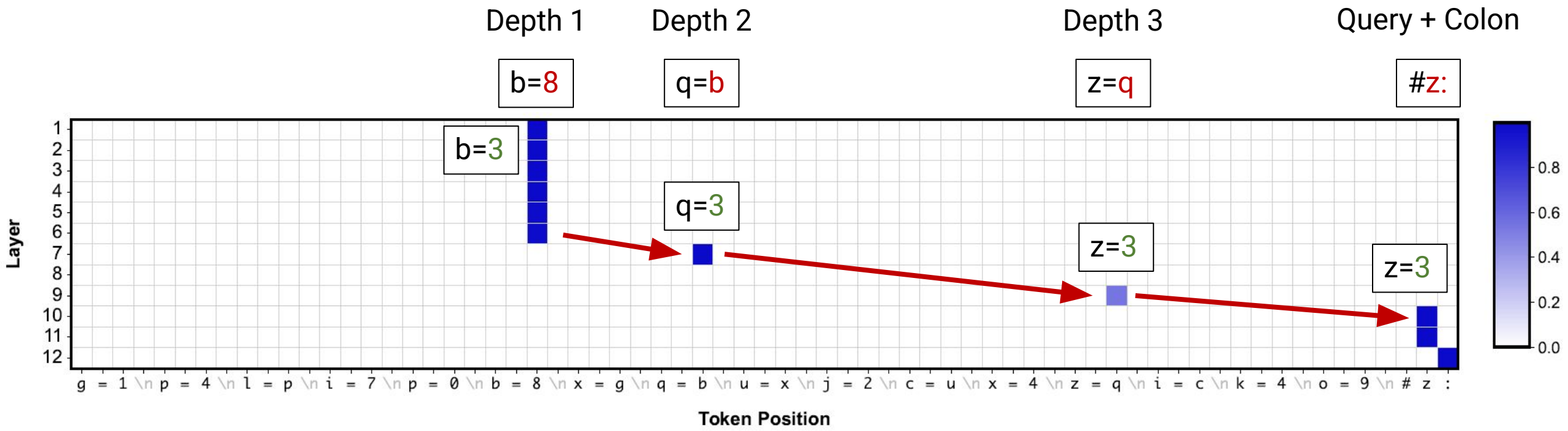


Patching the residual stream per layer



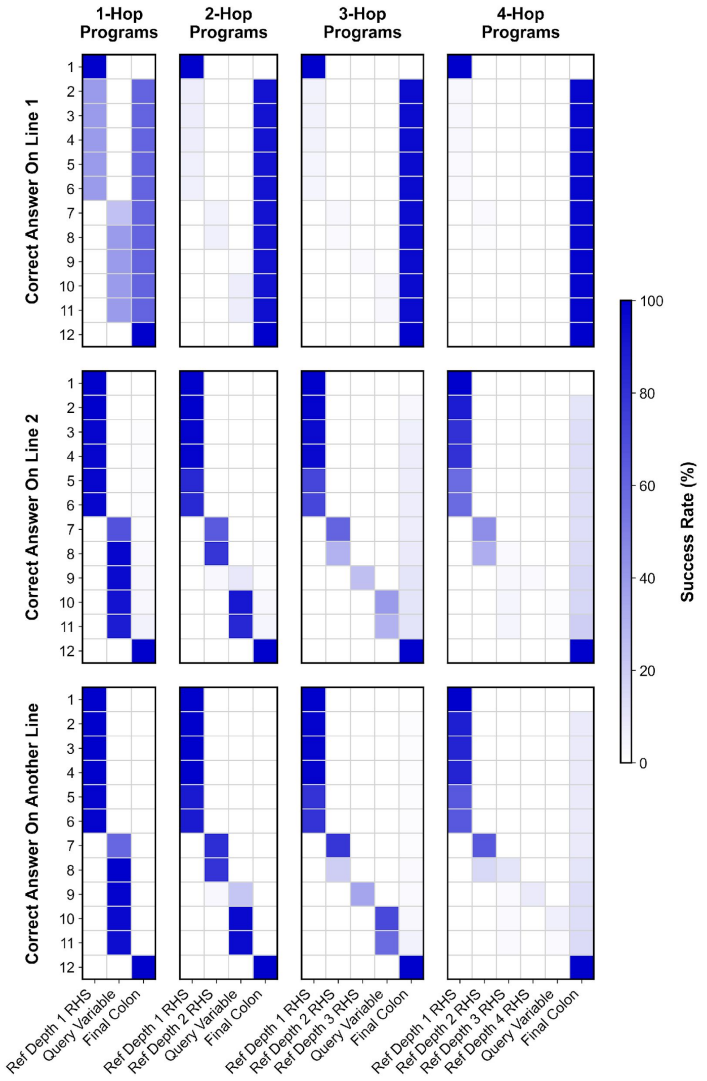


Patching the residual stream per layer



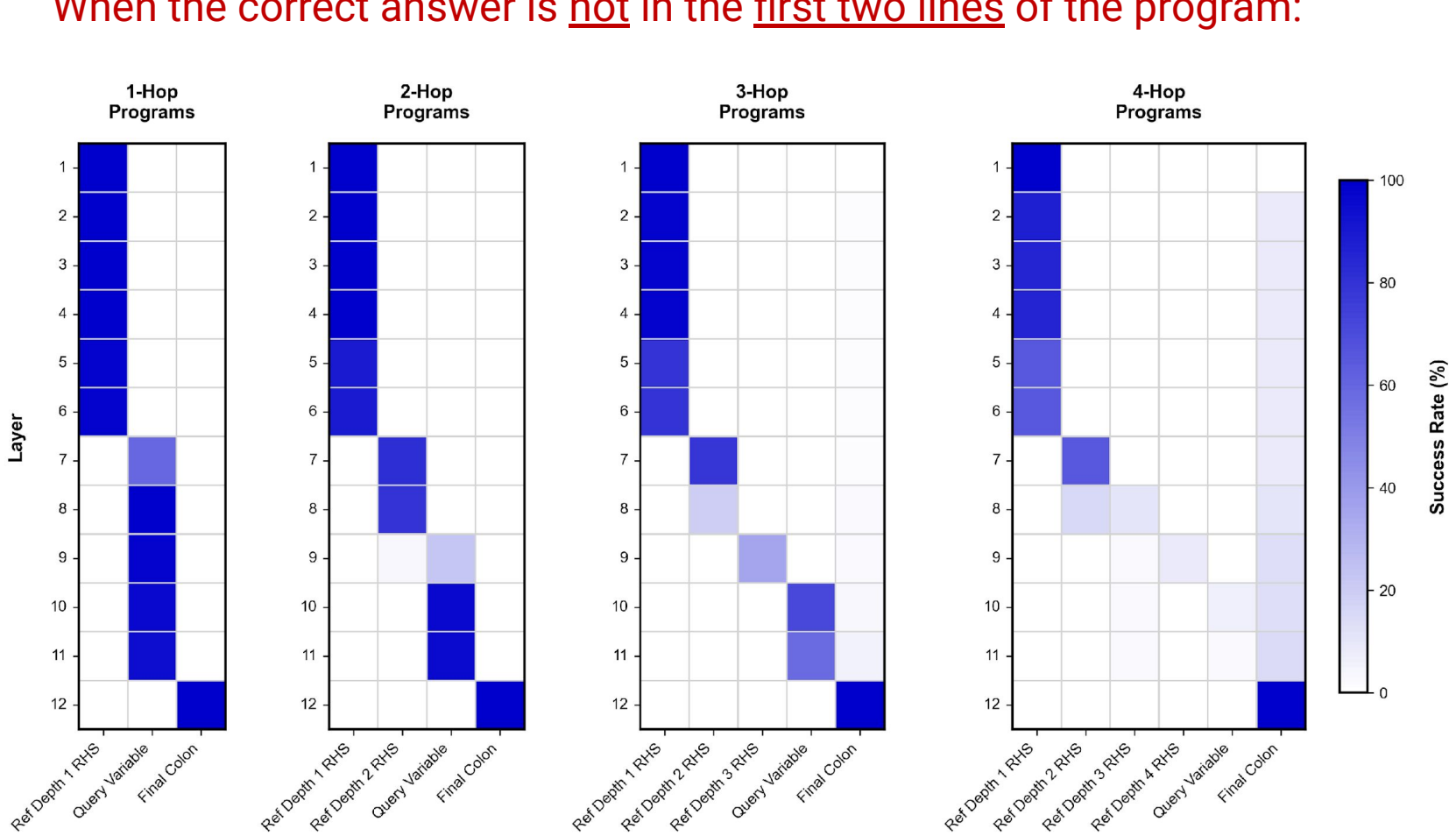
We focus on meaningful token positions (RHS at Ref Depth 1-4, Query, Colon) to aggregate patching results across programs.

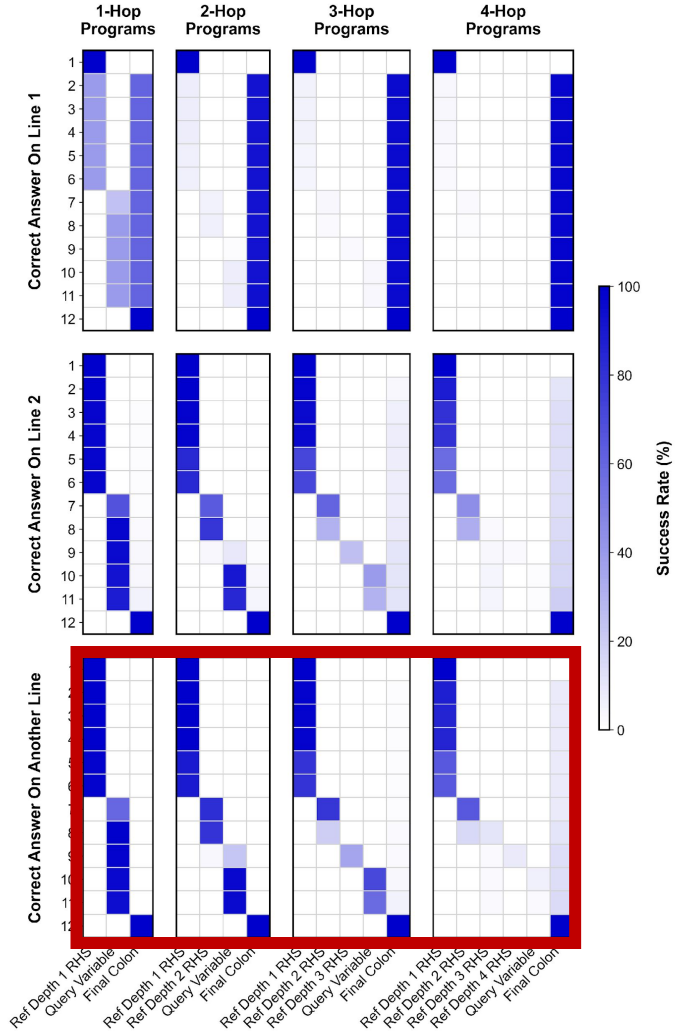
Patching the residual stream per layer



Patching the residual stream per layer

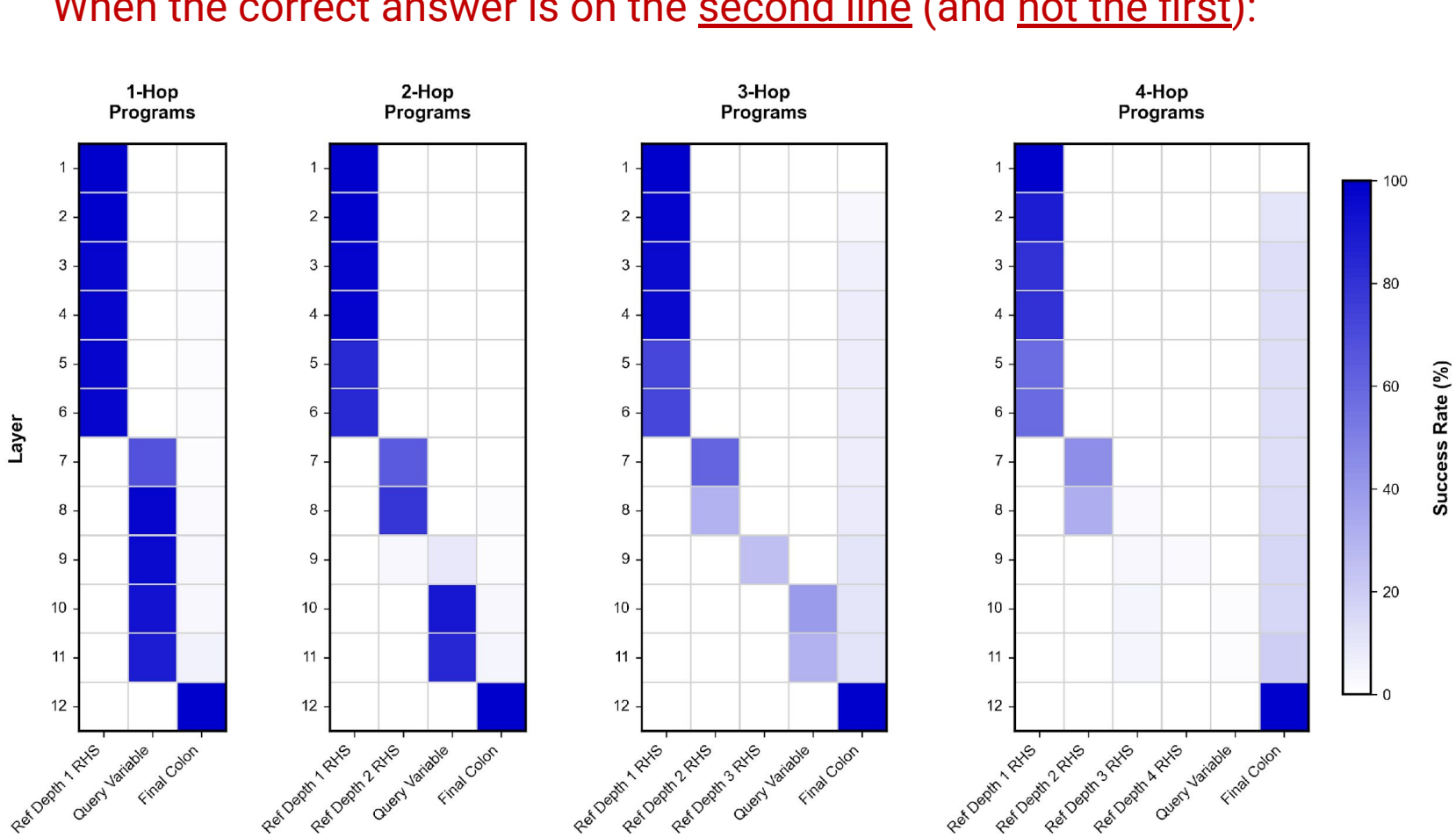
When the correct answer is not in the first two lines of the program:

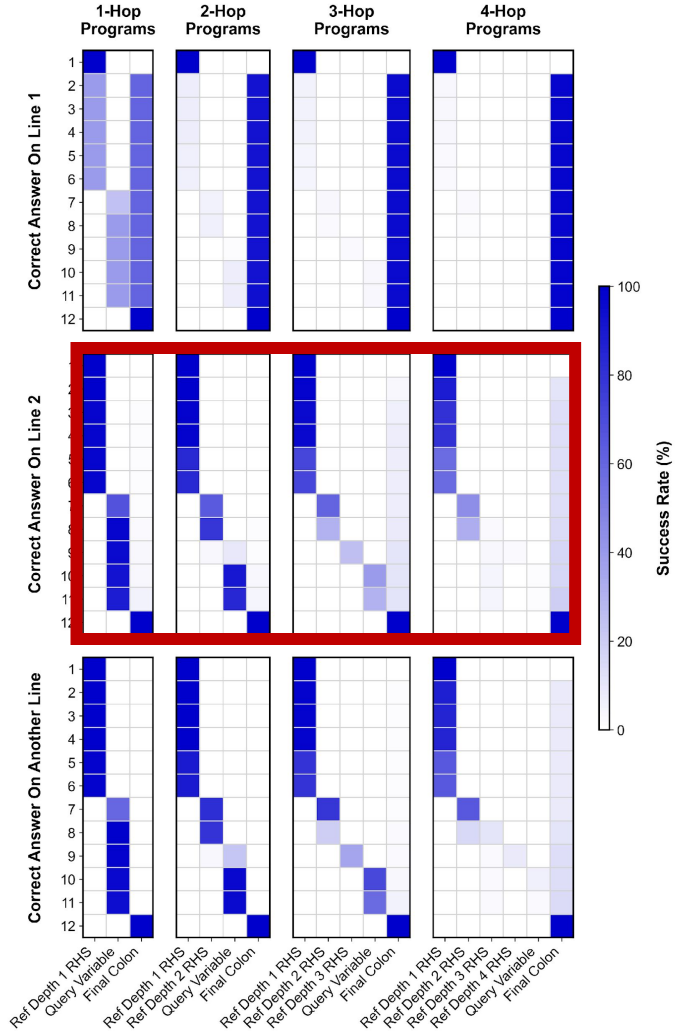




Patching the residual stream per layer

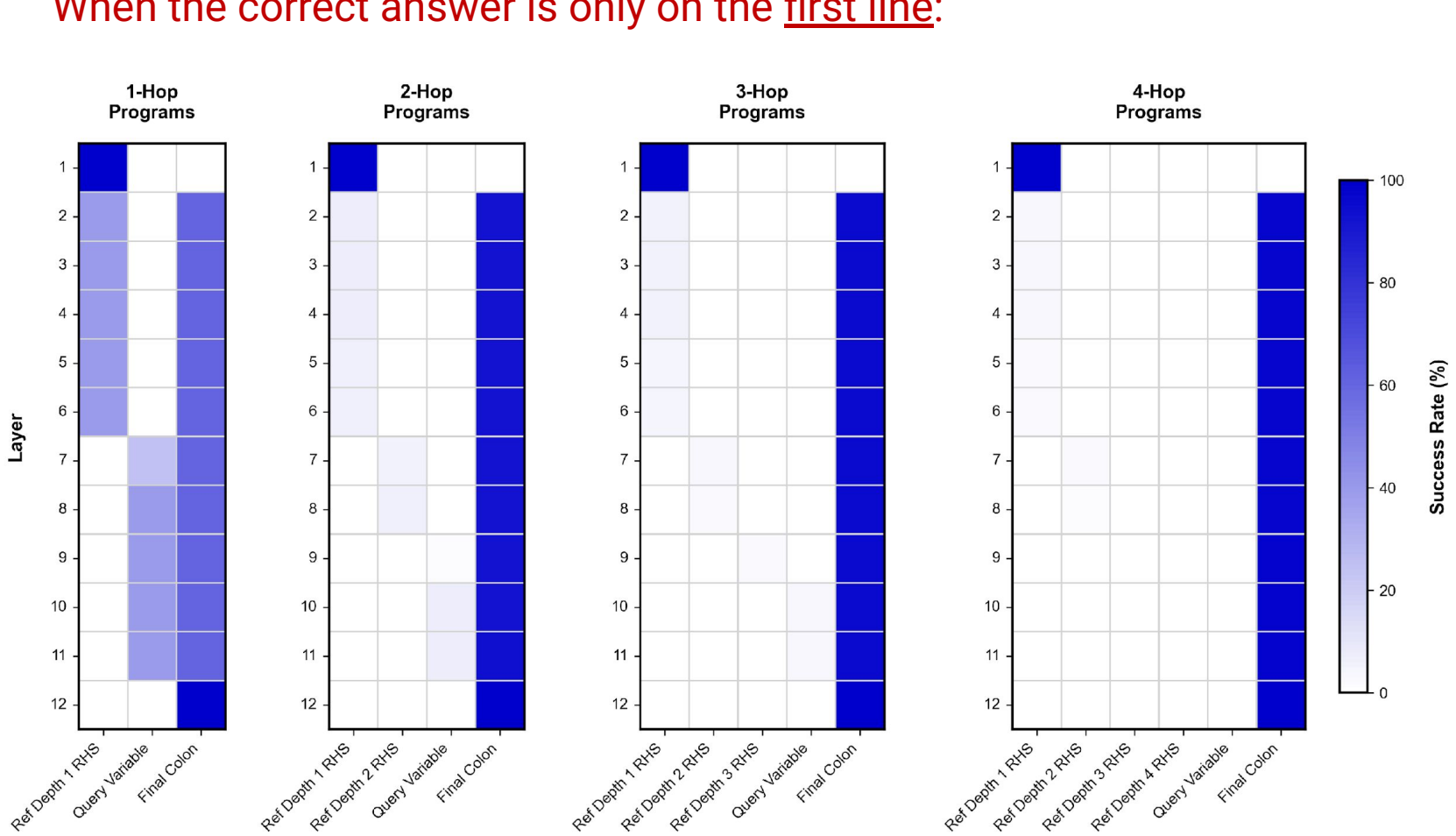
When the correct answer is on the second line (and not the first):

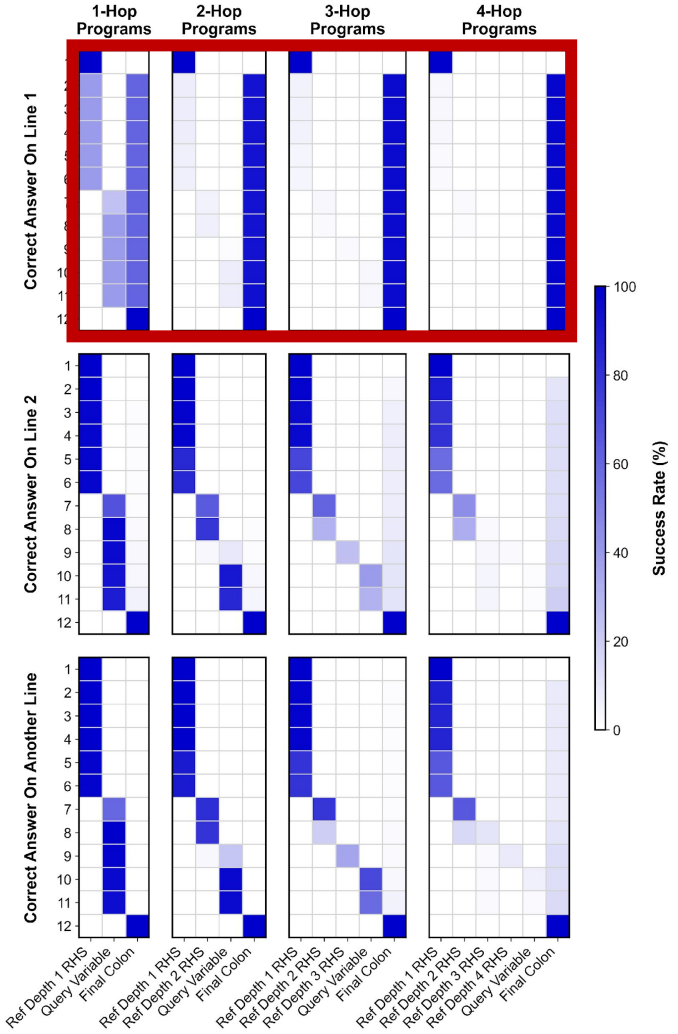




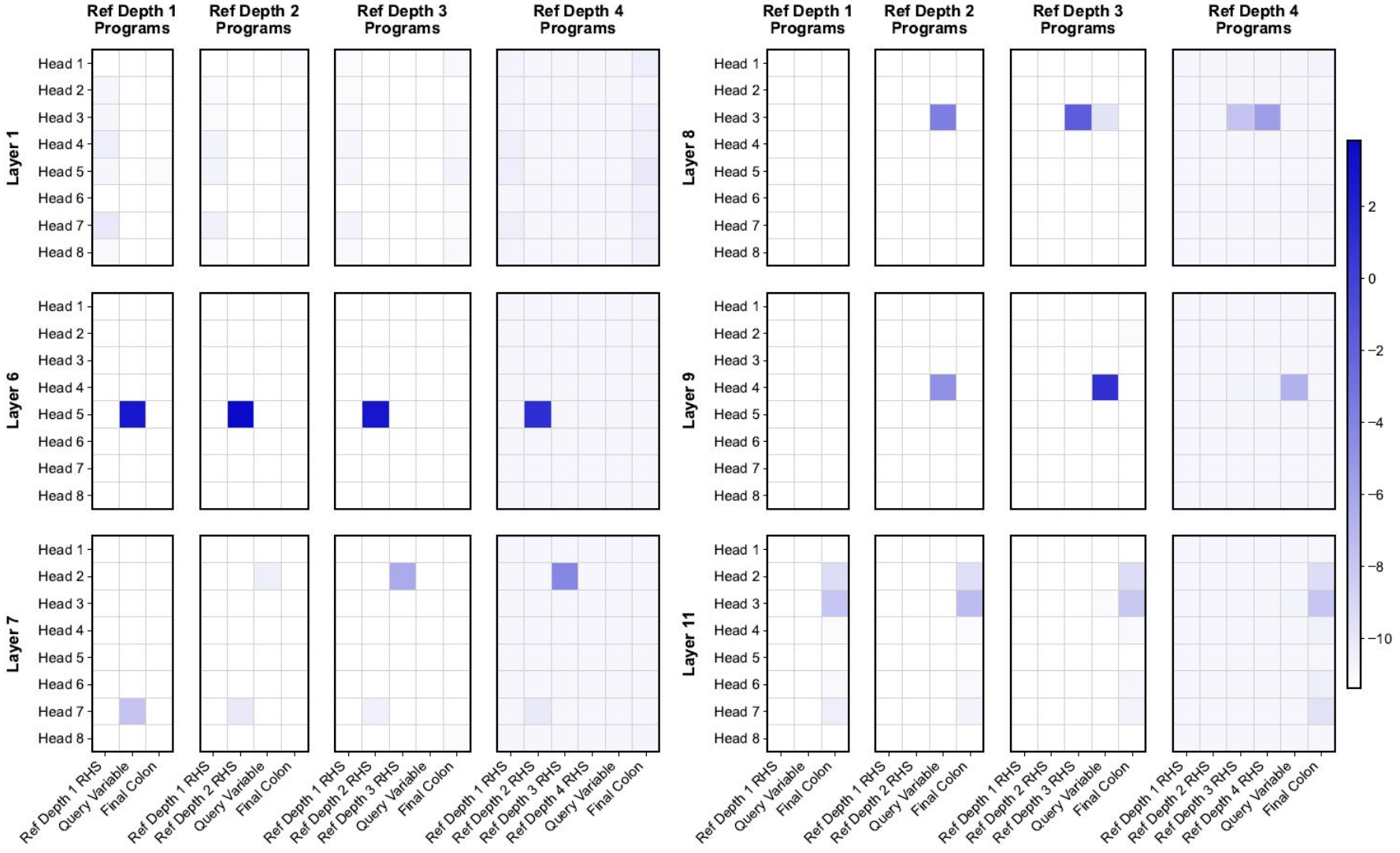
Patching the residual stream per layer

When the correct answer is only on the first line:





Patching the output of attention heads

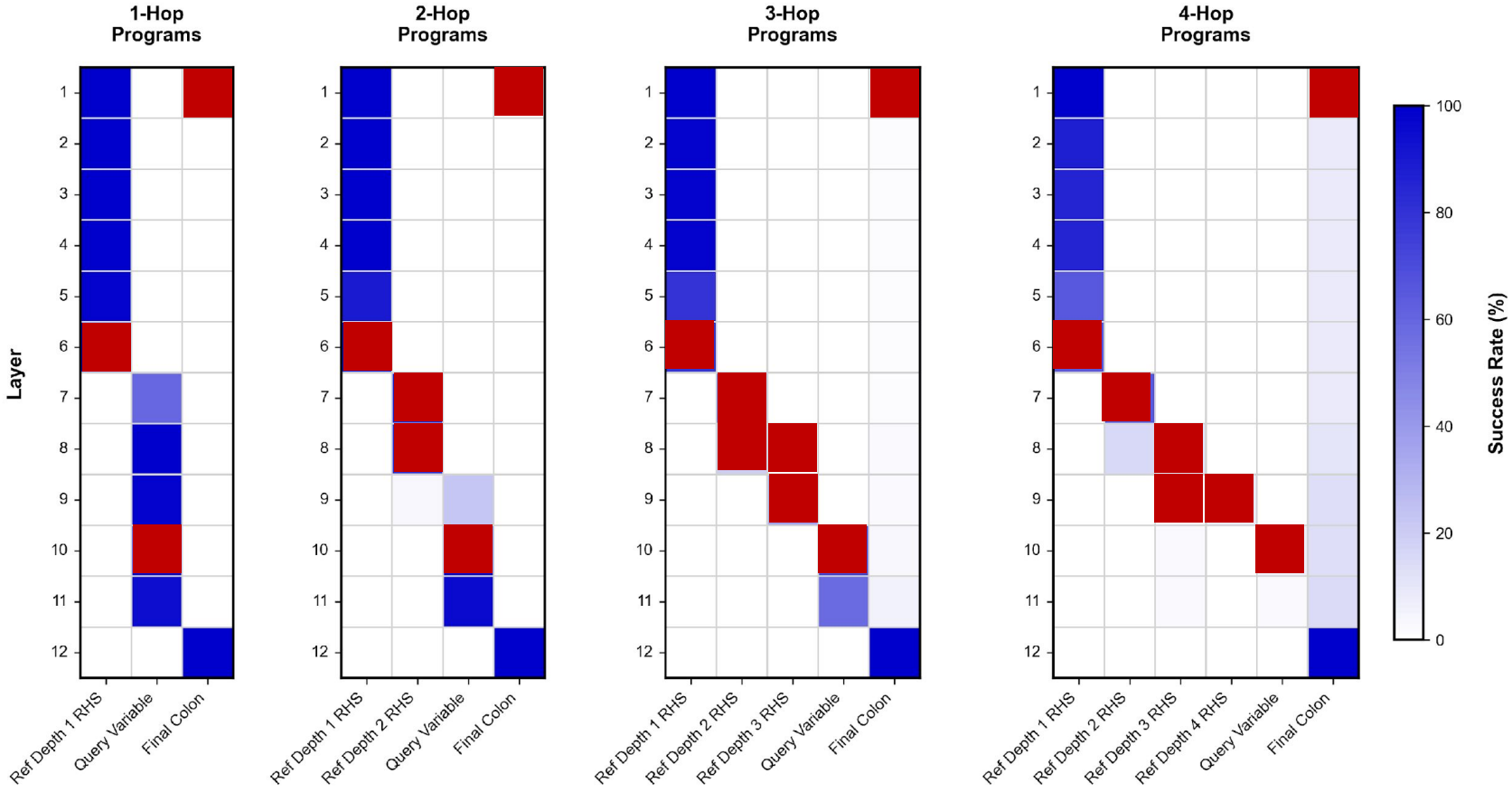


Tracing the developmental trajectory

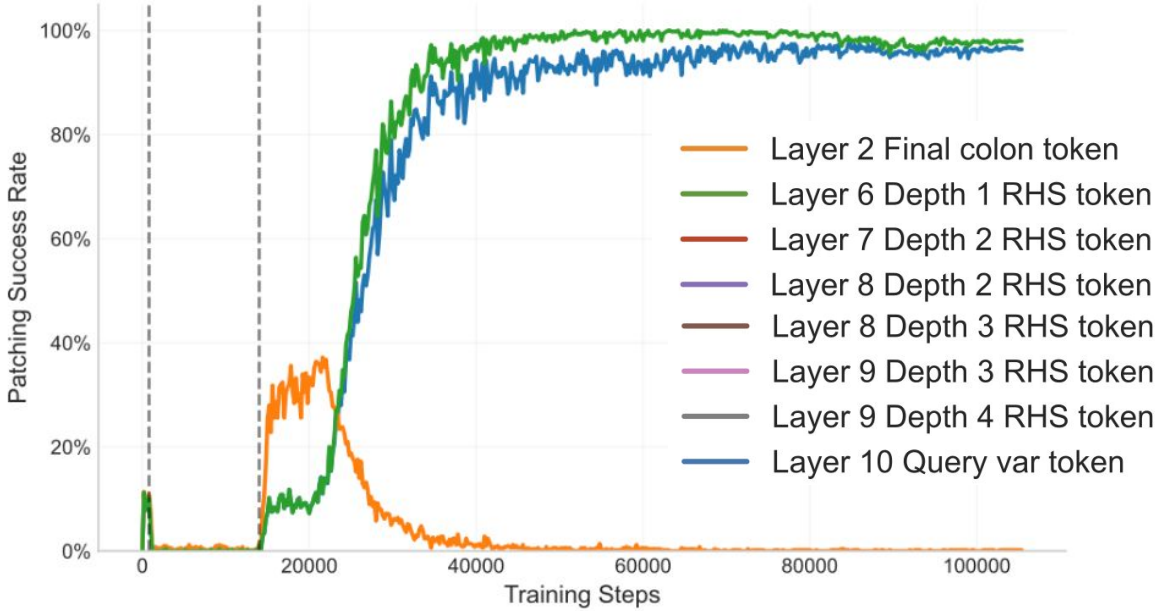
We tracked the evolution of patching success on the residual stream at key (layer, token) positions across training steps:

Layer 1 | Final colon

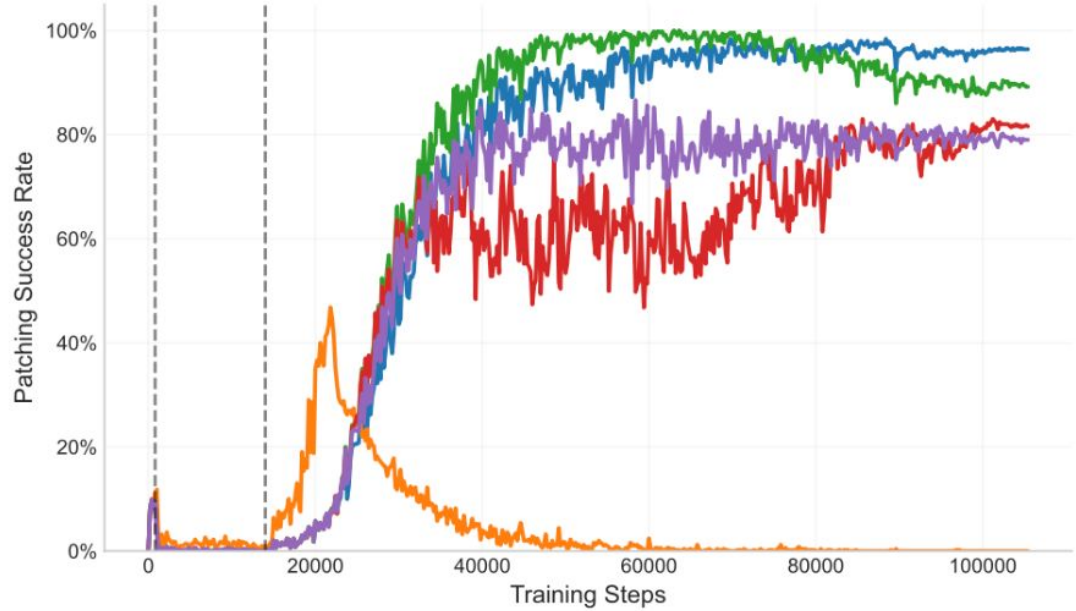
Layer 6 | Ref Depth 1
Layer 7 | Ref Depth 2
Layer 8 | Ref Depth 2 & 3
Layer 9 | Ref Depth 3 & 4
Layer 8 | Query



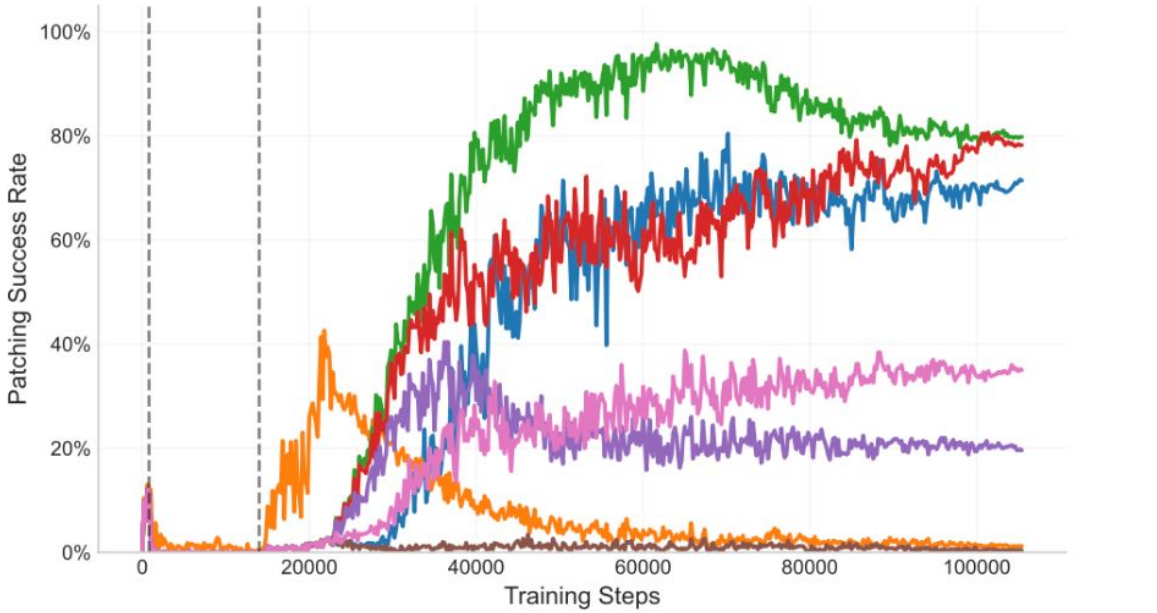
(a) Programs Where Correct Answer is on Line > 2 (Neither 1 Nor 2) — 1-Hop Programs



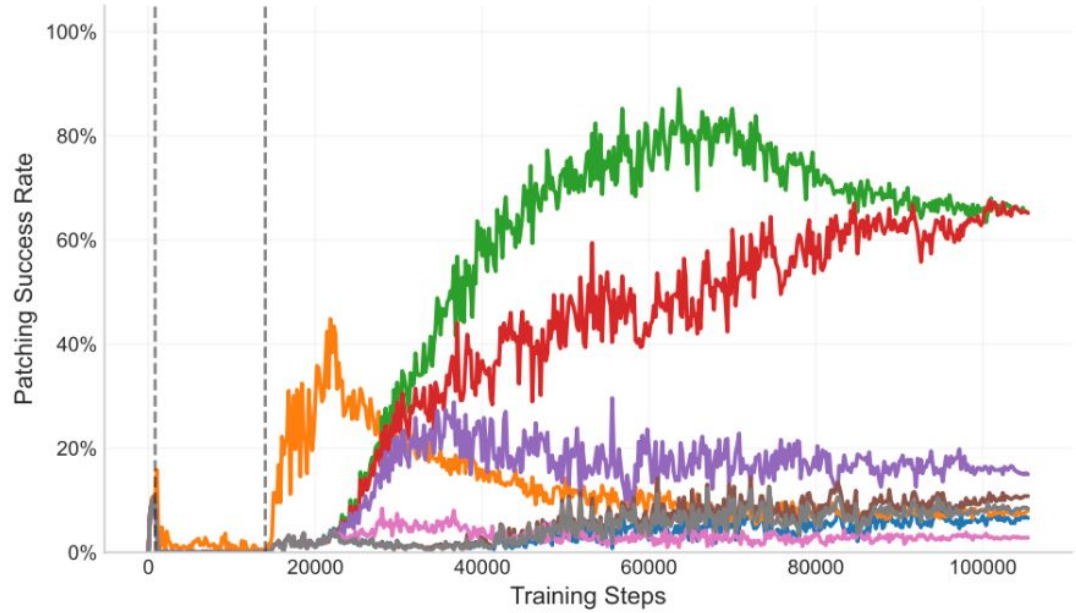
(b) Programs Where Correct Answer is on Line > 2 (Neither 1 Nor 2) — 2-Hop Programs

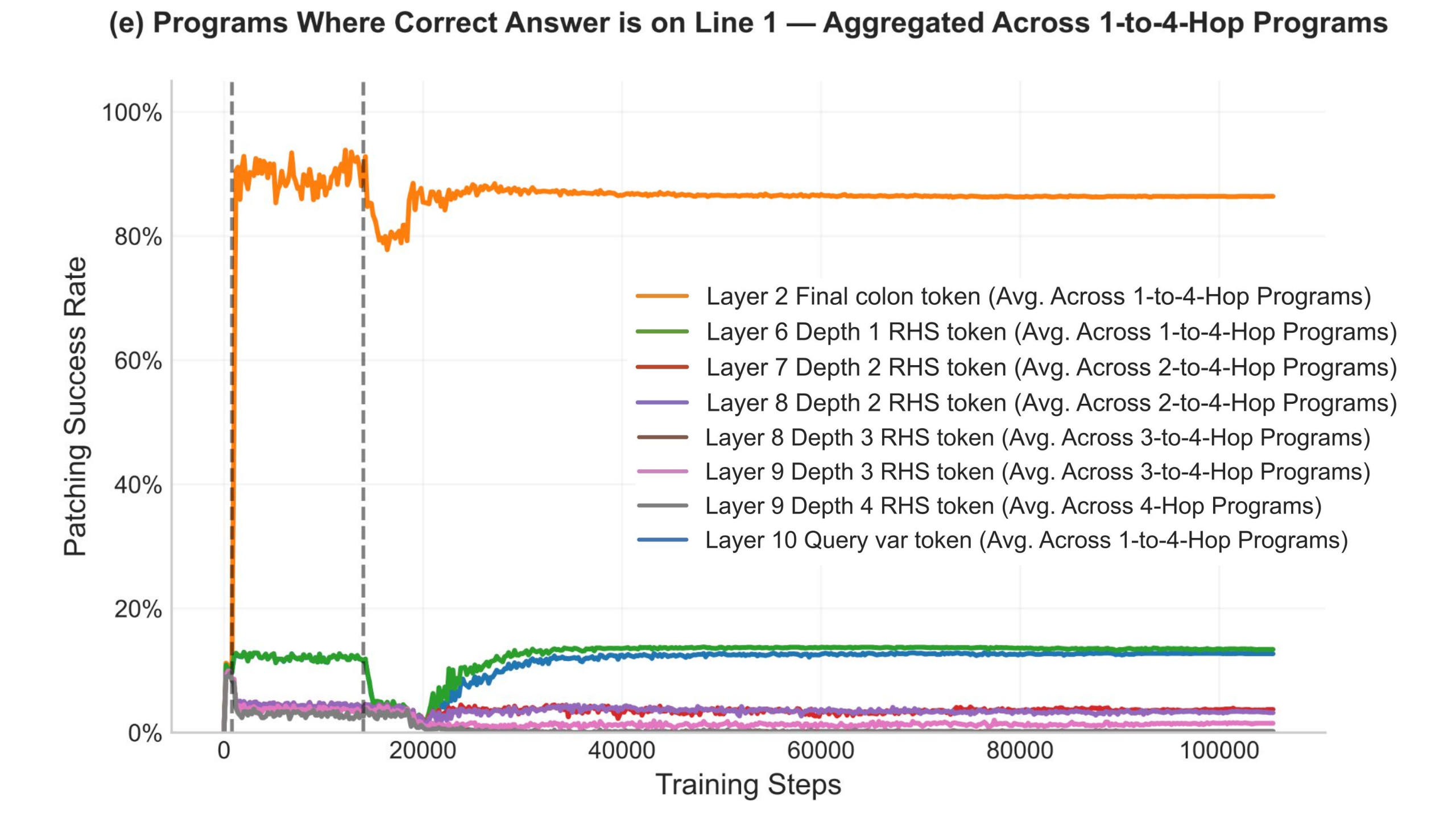


(c) Programs Where Correct Answer is on Line > 2 (Neither 1 Nor 2) — 3-Hop Programs

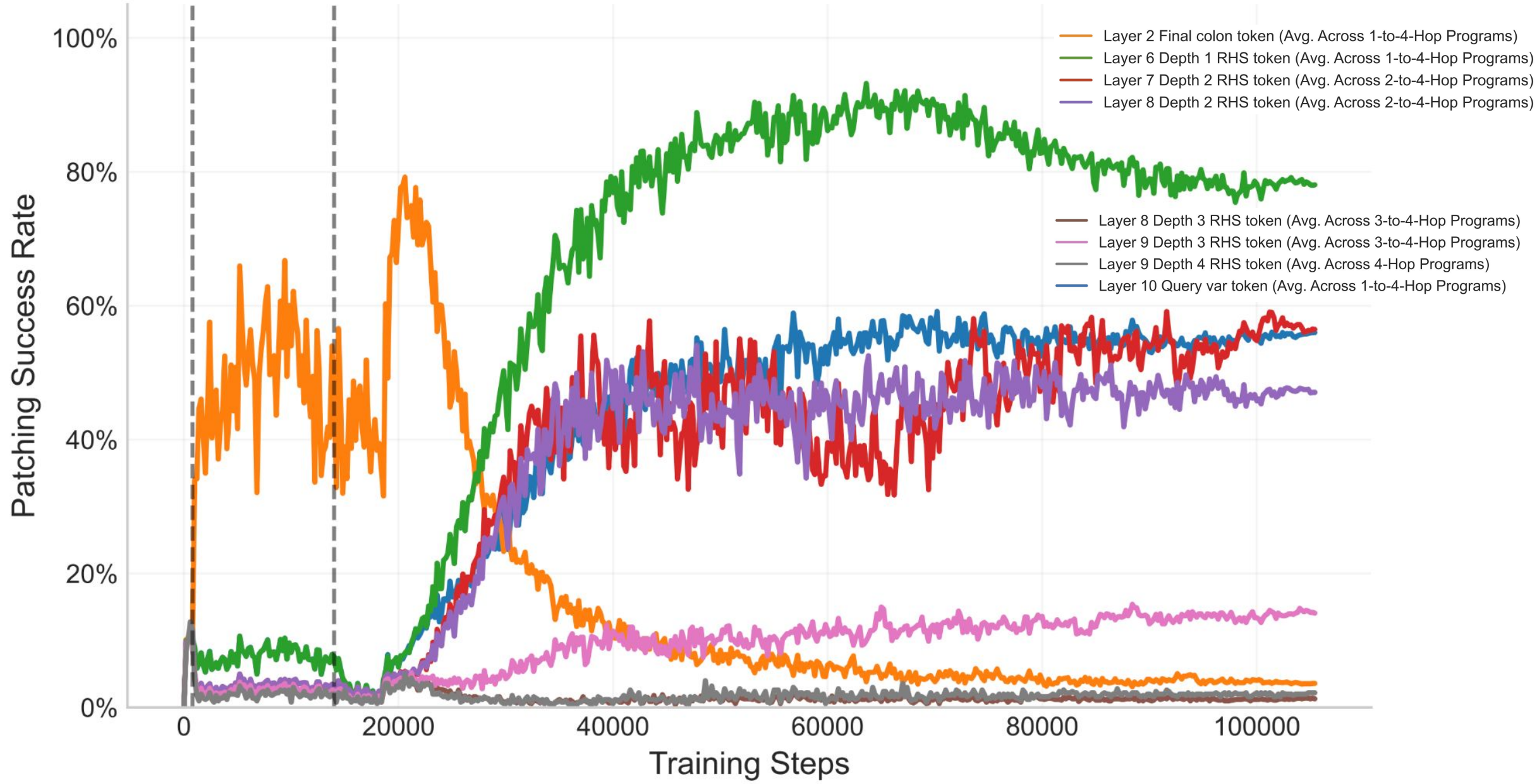


(d) Programs Where Correct Answer is on Line > 2 (Neither 1 Nor 2) — 4-Hop Programs

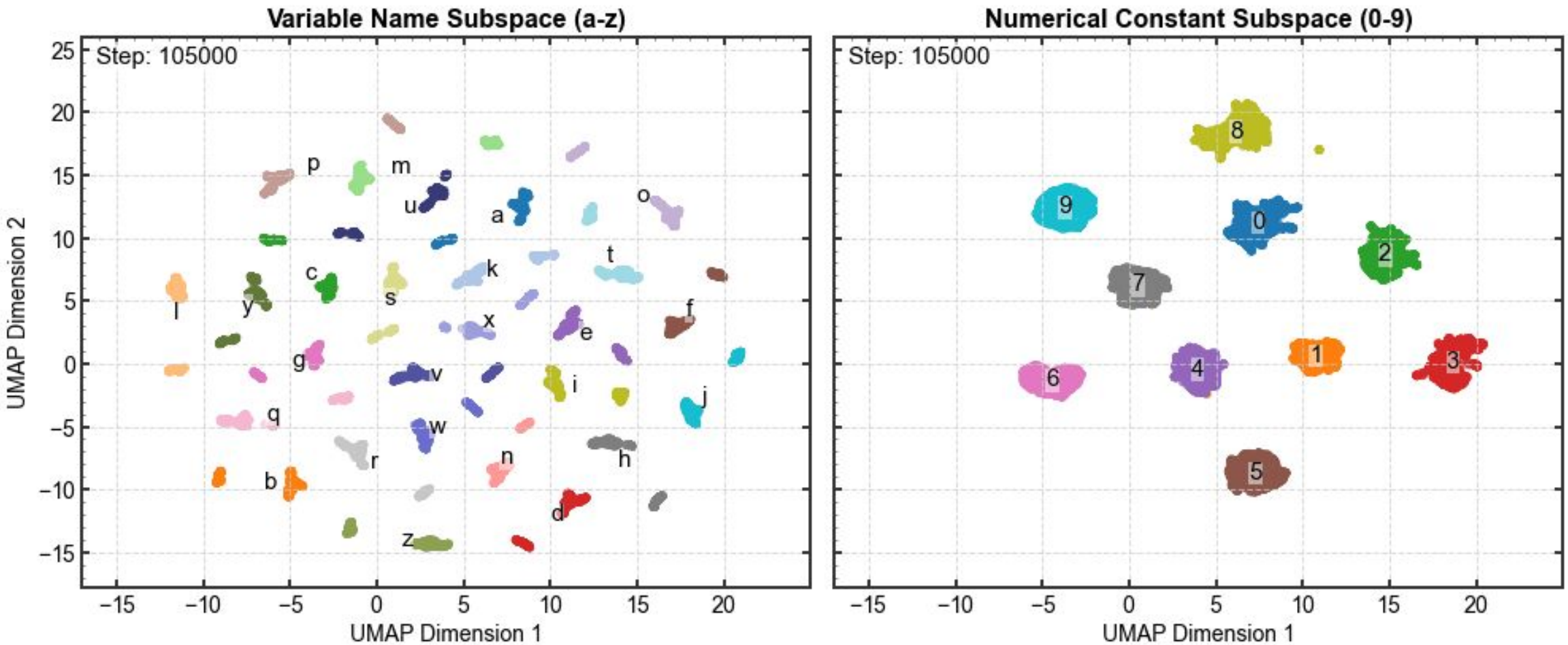



(e) Programs Where Correct Answer is on Line 1 — Aggregated Across 1-to-4-Hop Programs
Patching Success Rate
100%
80%
60%
40%
20%
0%
0
20000
40000
60000
80000
100000
Training Steps
Layer 2 Final colon token (Avg. Across 1-to-4-Hop Programs)
Layer 6 Depth 1 RHS token (Avg. Across 1-to-4-Hop Programs)
Layer 7 Depth 2 RHS token (Avg. Across 2-to-4-Hop Programs)
Layer 8 Depth 2 RHS token (Avg. Across 2-to-4-Hop Programs)
Layer 8 Depth 3 RHS token (Avg. Across 3-to-4-Hop Programs)
Layer 9 Depth 3 RHS token (Avg. Across 3-to-4-Hop Programs)
Layer 9 Depth 4 RHS token (Avg. Across 4-Hop Programs)
Layer 10 Query var token (Avg. Across 1-to-4-Hop Programs)

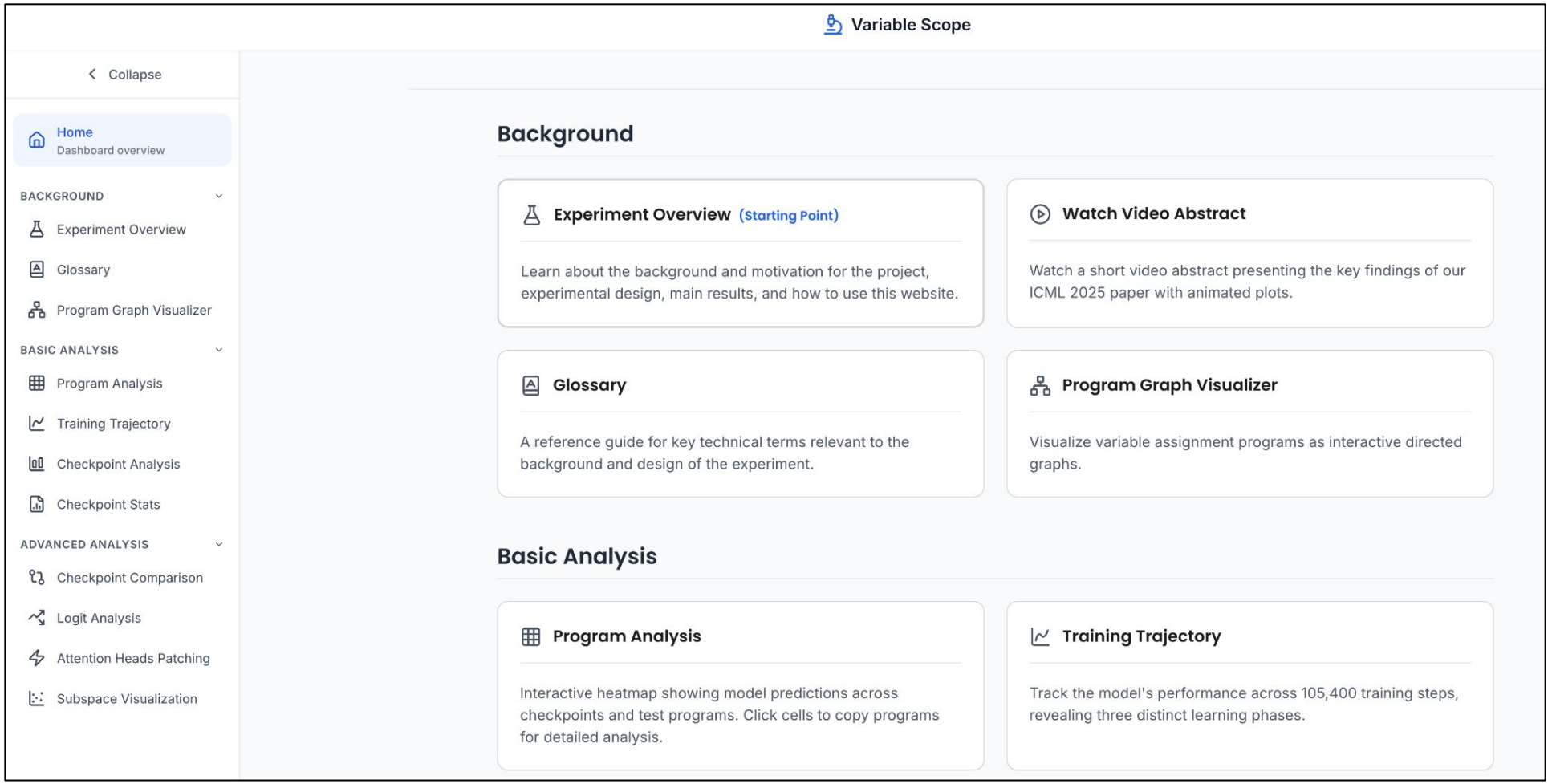
(f) Programs Where Correct Answer is on Line 2 (But Not 1) — Aggregated Across 1-to-4-Hop Programs



Two subspaces



Introducing: Variable Scope



variablescope.org

Questions