Transformers (can) learn to compose with independent circuits that “snap” together.

Assign and Add: A Mechanistic Study of Compositional Arithmetic

Background: We train a 2-layer transformer on variable assignment and modular addition and ask how it gneralizes and composes the two operations.

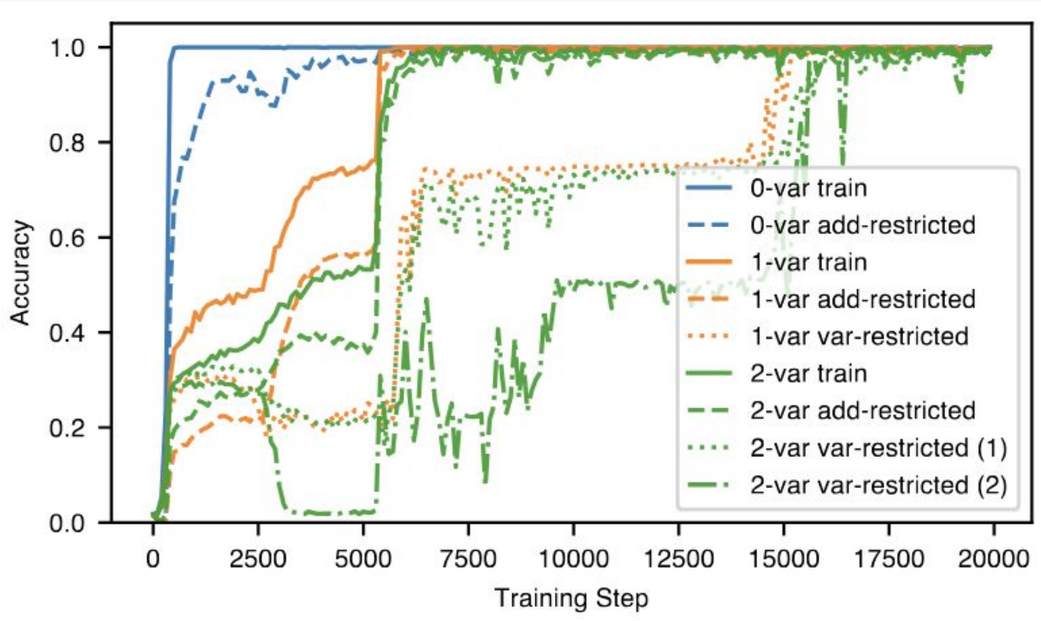
Example Sequences

PAD	PAD	PAD	PAD	c	1	PAD	f	17	a	42	PAD	+	f	19	=	36
b	2	PAD	g	08	PAD	c	53	PAD	PAD	e	14	+	17	32	=	49
a	3	c	58	PAD	d	19	f	06	b	47	PAD	+	b	f	=	53

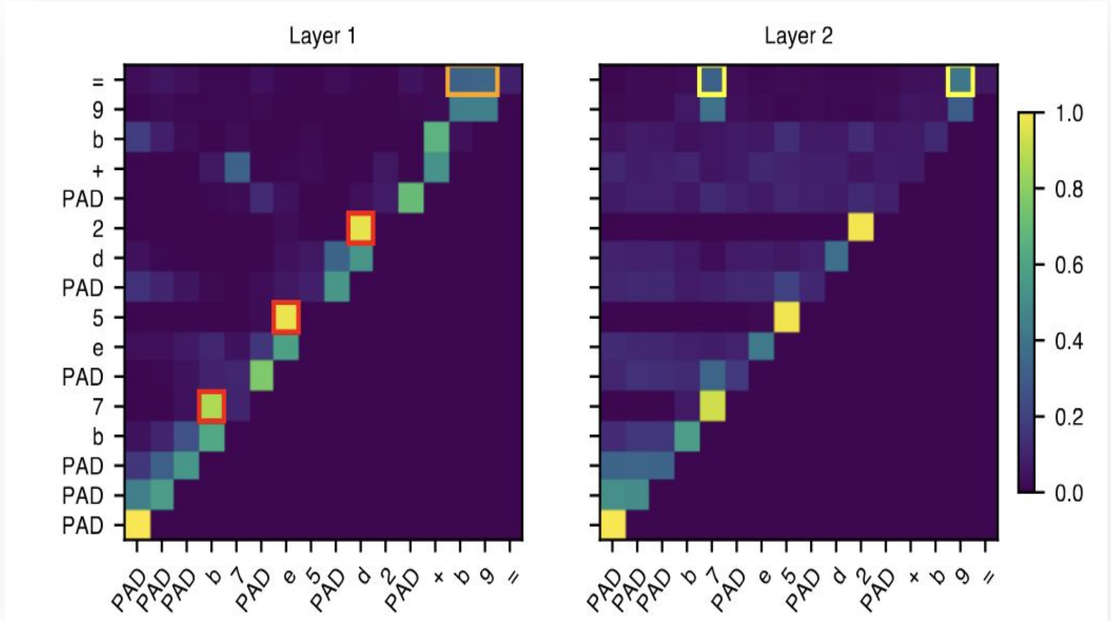
Result 1: The addition and variable assignment circuits are learned in separate “phases”.

Result 2: The variable assignment circuit “feeds” into the already-learned addition circuit.

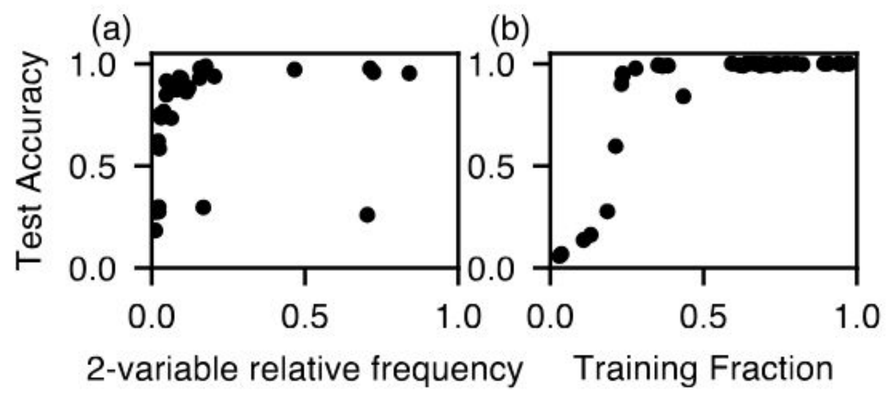
1



2



3



Result 3: By composing these two operations, the model can generalize to unseen distributions,

Limitation: Although our study provides insights into a controlled setting, we only consider a task and model of limited complexity in order to ease analysis.

PAPER

Brady Exoo, Alberto Bietti, John Sous

Mechanistic Interpretability Workshop, ICML 2026