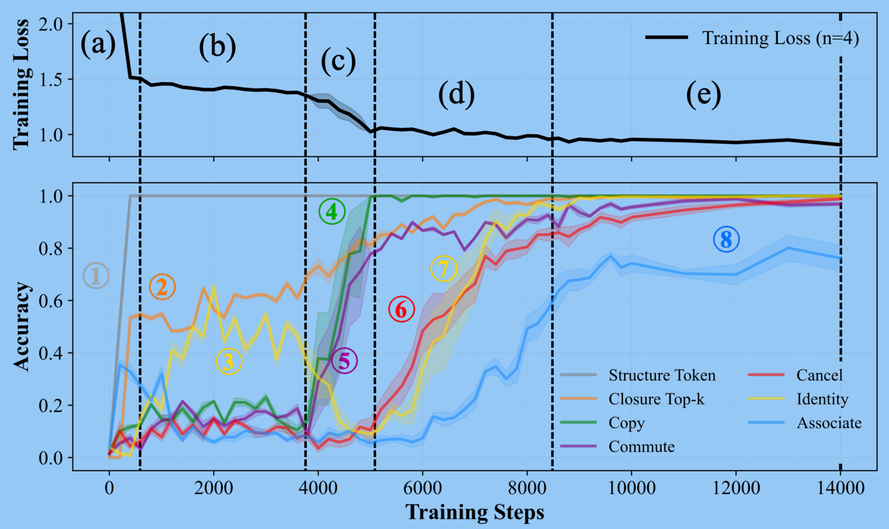
Transformers infer **structure** from **variables**



In-Context Algebra

Eric Todd, Jannik Brinkmann, Rohit Gandikota, David Bau

Website: algebra.baulab.info

Northeastern University

$12=3$ $43=2$ $01=1$ $\cdots$ $14=0$ $21=0$ $20=2$
$pn=i, ka=b, jd=d, \cdots, lk=g, ed=j, ej=$ e

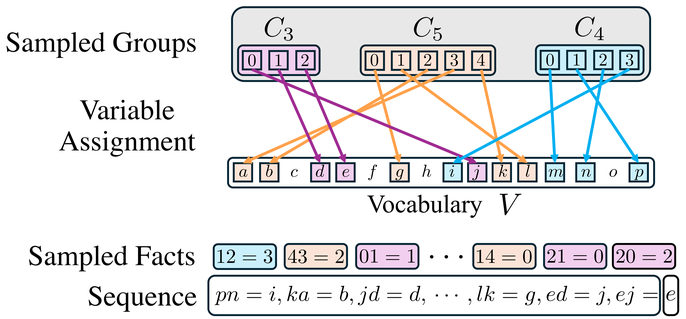
$30=3$ $41=5$ $60=6$ $\cdots$ $24=0$ $21=7$ $13=4$
$bd=b, ai=c, gf=g, \cdots, eh=d, li=k, pb=$ h

$13=4$ $20=2$ $34=0$ $\cdots$ $65=4$ $23=5$ $66=5$
$fp=k, ae=a, pk=e, \cdots, lc=k, ap=c, ll=$ h

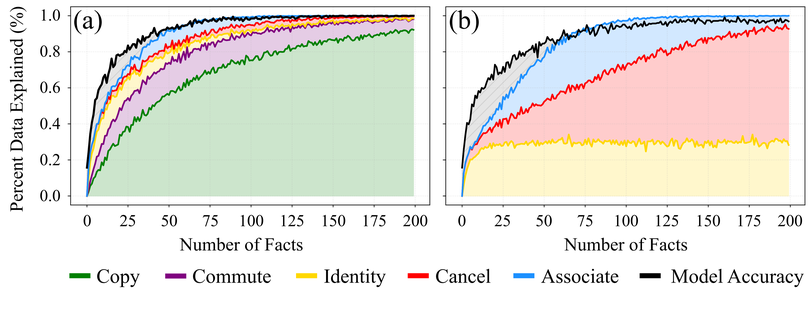
Variables allow us to reason about objects with unknown meaning.

How do transformers **reason** about variables **in-context**?

(1) We design a **contextual reasoning** task where **interactions** define tokens

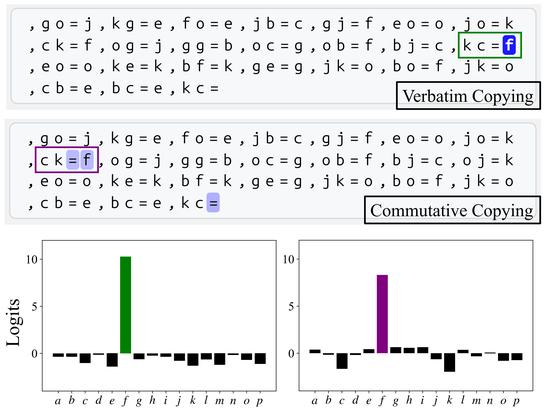


(2) We **isolate** parallel reasoning strategies **using targeted data** distributions

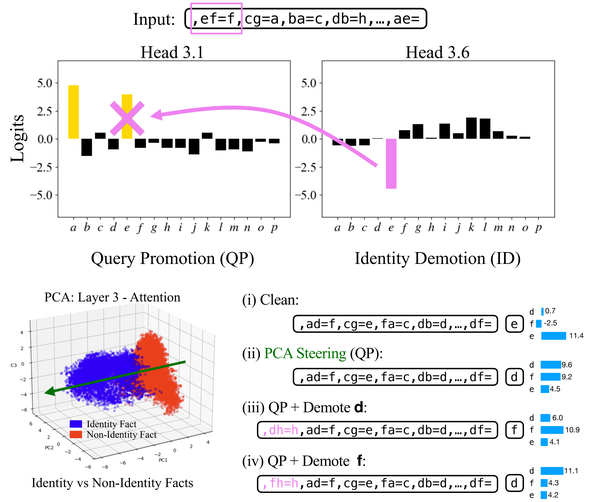


(3) Transformers develop discrete symbolic strategies to reason about variables

Commutative Copying



Identity Recognition



Cancellation