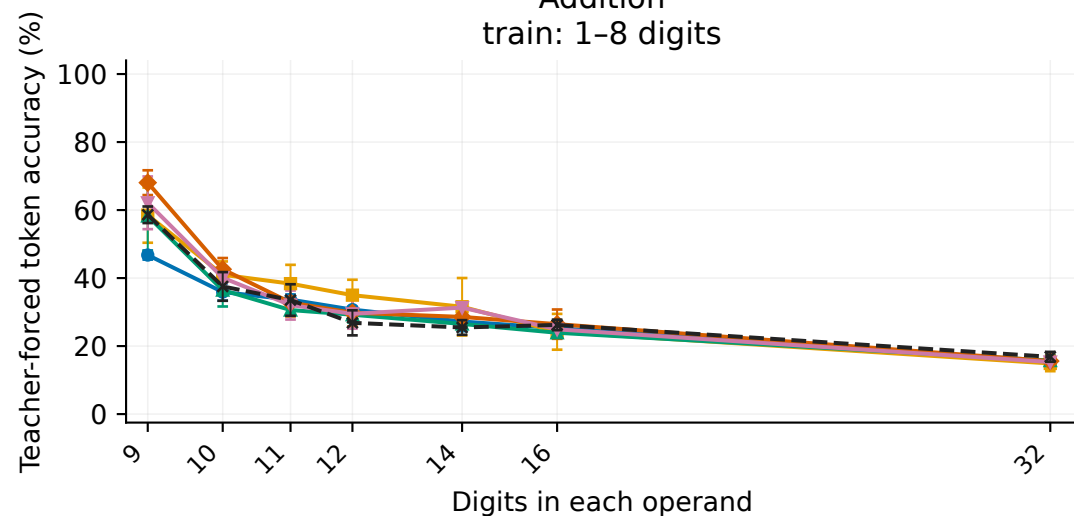


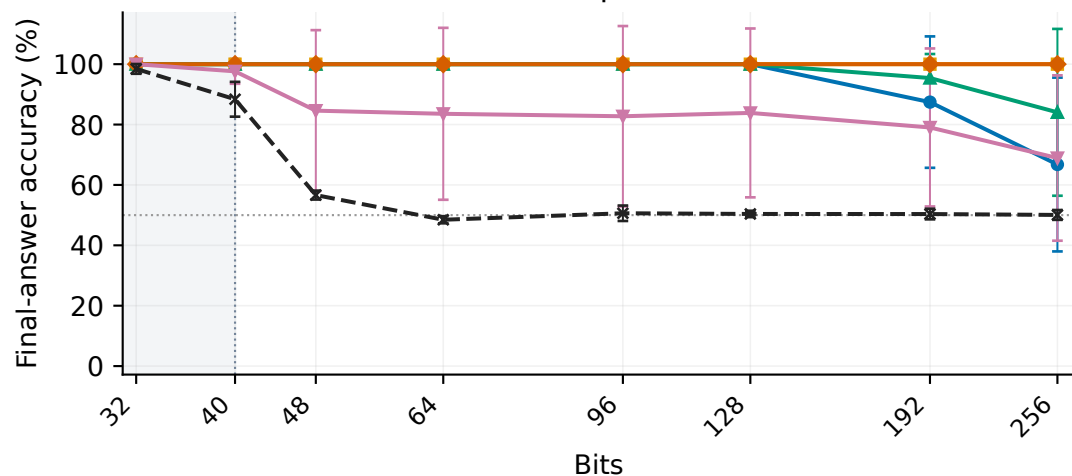
Length generalization · 8 layers · seeds 42, 43, 44

RLT-1 4+4 RLT-1 6+2 RLT-1 8+0
 RLT-1 5+3 RLT-1 7+1 Transformer 8

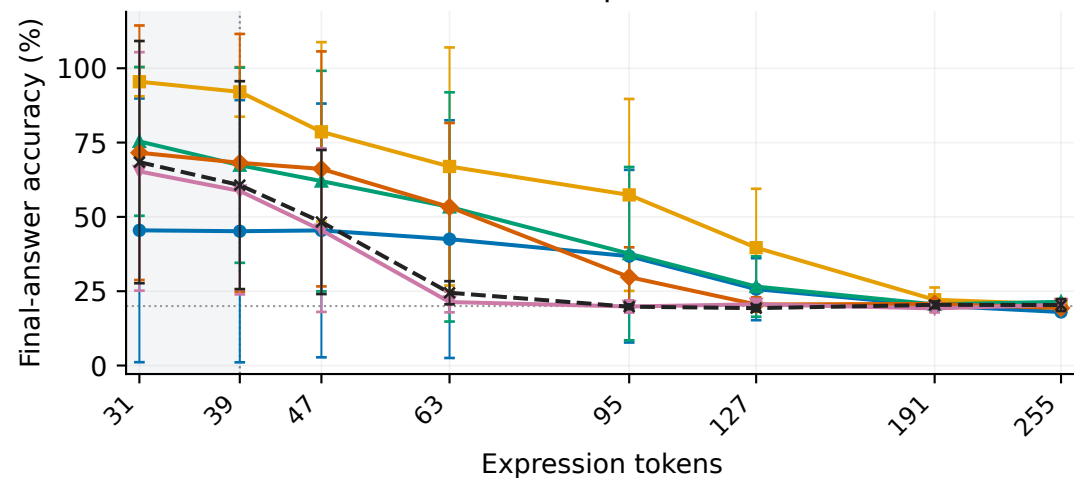
Addition
train: 1-8 digits



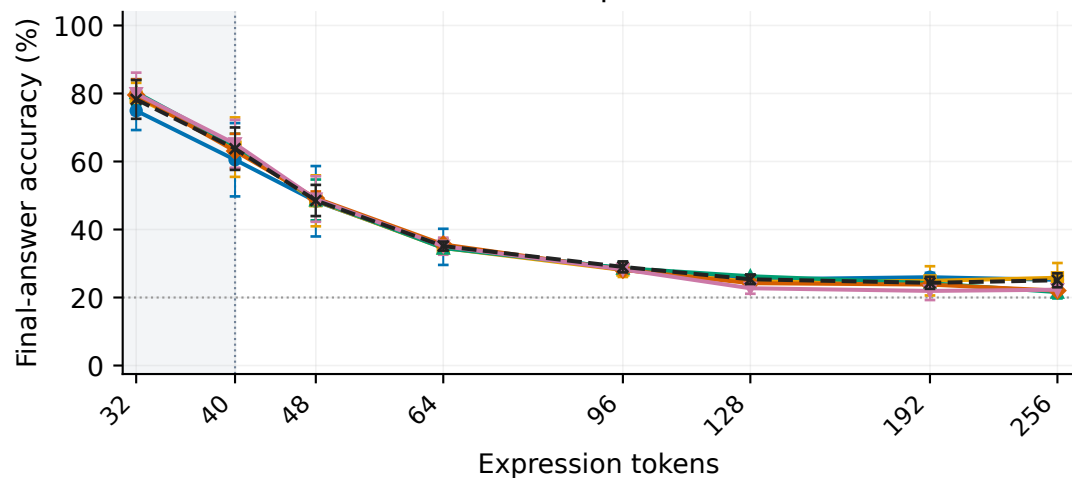
Parity
train: up to 40



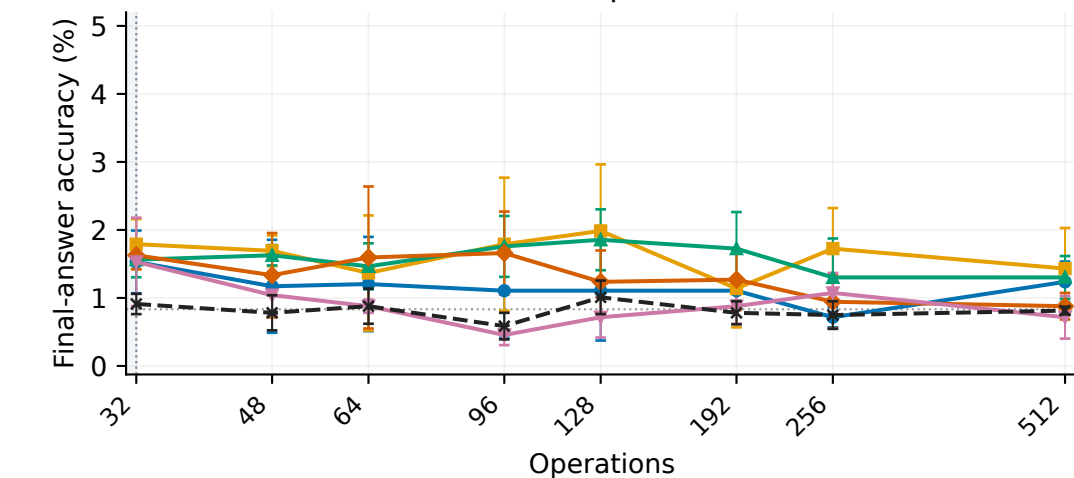
Mod 5: no brackets
train: up to 39



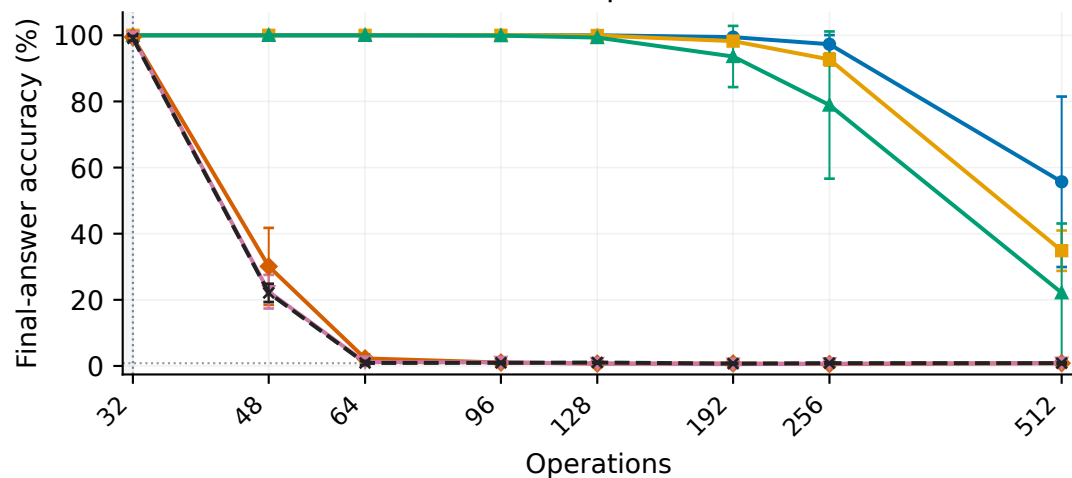
Mod 5: with brackets
train: up to 40



S_5 : standard
train: 32 operations



S_5 : swaps
train: 32 operations



Mean $\pm$ sample SD (n=3) · best ID-validation-loss checkpoint per run
 256 addition pairs / 1,024 formal sequences per length · shared test examples