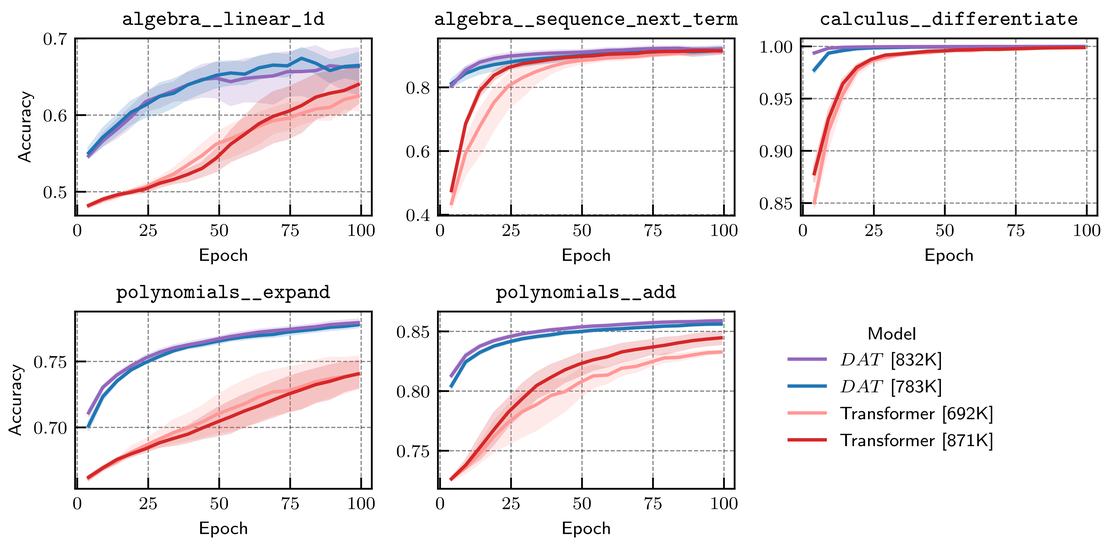
Disentangling and Integrating Relational and Sensory Information in Transformer Architectures

Relational games

We computed learning curves on relational games, comparing *DAT* against multiple Transformer baselines of varying sizes and architectural hyperparameters (e.g., # of heads). We also computed learning curves on relational games, comparing *DAT* against PrediNet, CoRelNet, Abstractor, and Transformer baselines.

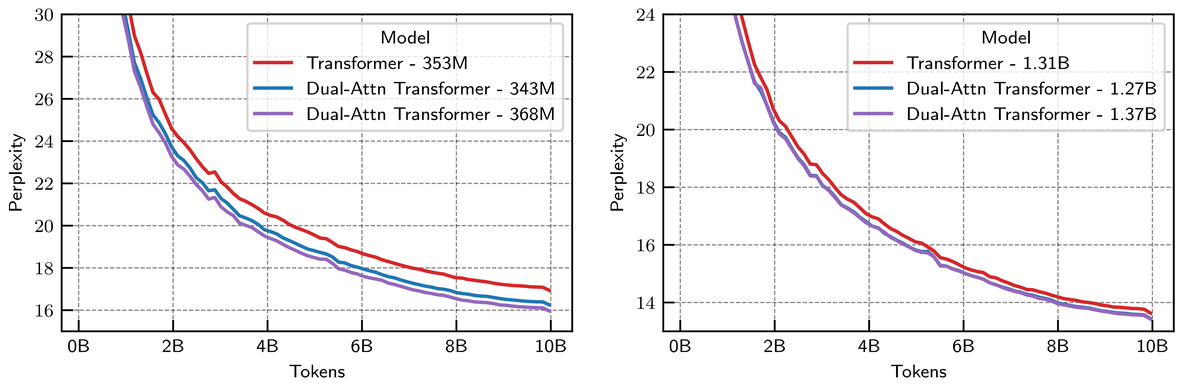
Mathematics processing

Validation accuracy over the course of training for seq2seq mathematical problem-solving:



We also ran sequence-to-sequence symbolic mathematical processing, comparing to a Transformer at multiple scales, with *DAT* models of model dimension 128 and Transformer models of model dimension 144, with three models each with 2, 3, or 4 layers. Superiority of *DAT* persists across all depths and model sizes. Figures included in revised paper.

Language modeling at larger scale



Plots show perplexity curves on language modeling with the Fineweb dataset. The x-axis indicates the number of tokens and the y-axis is the validation perplexity. *DAT* learns faster and achieves smaller perplexity at multiple model size scales.

Model	Param count	# Tokens	d_{model}	n_{layers}	n_h^{sa}	n_h^{ra}	d_r	n_{kv}^h	Perplexity ↓
Transformer	353M	10B	1024	24	16	-	-	-	16.94
DAT	343M	10B	1024	24	8	8	32	4	16.26
DAT	368M	10B	1024	24	8	8	32	8	15.97
Transformer	1.31B	10B	2048	24	32	-	-	-	13.63
DAT	1.27B	10B	2048	24	16	16	64	8	13.44
DAT	1.37B	10B	2048	24	16	16	64	16	13.43