Table 1: Performance comparison of A2P Scaffolding against baseline methods on both datasets. Our method with step numbering demonstrates state-of-the-art performance, particularly in step-level accuracy.

	Algorithm-Generated (126 samples)				**Hand-Crafted (58 samples)**			
Method	**Agent Accuracy (%)**		**Step Accuracy (%)**		**Agent Accuracy (%)**		**Step Accuracy (%)**	
	Value	Gain	Value	Gain	Value	Gain	Value	Gain
A2P (Ours)	**65.40**	–	**47.46**	–	**58.62**	–	**29.31**	–
Baselines								
all_at_once	63.49	-1.91	16.67	-30.79	27.59	-31.03	12.07	-17.24
step_by_step	49.21	-16.19	27.78	-19.68	53.45	-5.17	18.97	-10.34
binary_search	46.83	-18.57	28.57	-18.89	44.83	-13.79	13.79	-15.52