Table 10: **Extended results for experiment Q1.** We evaluated all models over 30 standard (*i.e.*, non-optimal) runs using different seeds and measured their label accuracy (Acc($\mathbf{Y}$)) and concept accuracy (Acc($\mathbf{C}$)).

	XOR					
	DPL		SL		LTN	
	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)
–	$80.6 \pm 16.9\%$	$22.0 \pm 7.1\%$	$89.3 \pm 15.9\%$	$23.7 \pm 7.0\%$	$83.9 \pm 14.5\%$	$20.5 \pm 9.0\%$
DIS	$82.1 \pm 24.0\%$	$68.8 \pm 41.9\%$	$84.2 \pm 21.6\%$	$82.5 \pm 35.0\%$	$93.3 \pm 17.0\%$	$88.8 \pm 28.8\%$

	MNIST-Addition					
	DPL		SL		LTN	
	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)	Acc($\mathbf{Y}$)	Acc($\mathbf{C}$)
–	$96.3 \pm 1.2\%$	$43.1 \pm 21.3\%$	$97.0 \pm 0.3\%$	$40.2 \pm 26.3\%$	$78.5 \pm 25.1\%$	$46.3 \pm 17.1\%$
DIS	$99.5 \pm 0.1\%$	$99.7 \pm 0.1\%$	$98.8 \pm 0.1\%$	$99.7 \pm 0.1\%$	$95.8 \pm 5.5\%$	$97.8 \pm 3.0\%$