Table 1: Aggregated correctness results for the different tools considering each scenario.

		User Interaction	Unit Test
Average Correctness	**Text**	17.20%	77.80%
	k-Tails	58.30%	61.10%
	2KDiff &nKDiff	97.80%	97.20%
p-value / δ	**Text**	$\leq$0.01 / 1.00 (large)	$\leq$0.01 / 0.86 (large)
	k-Tails	$\leq$0.01 / 0.85 (large)	$\leq$0.01 / 0.79 (large)