Challenge Information			Human Baseline Metrics				MLAB			OpenHands			AIDE (o1-preview)			Agent Laboratory `mle-solver` (ours)		
Challenge Title	Data Type	Min/Max?	Median Score	Bronze Medal	Silver Medal	Gold Medal	Score	Above Median	Medal Earned	Score	Above Median	Medal Earned	Score	Above Median	Medal Earned	Score	Above Median	Medal Earned
detect insults in commentary		Max ↑	0.778	0.791	0.823	0.833	0.749	✗		0.867	✓	🥇	**0.904**	✓	🥇	**0.839**	✓	🥇
dec 2021 tab playground		Max ↑	0.953	0.956	0.956	0.956	0.828	✗		0.957	✓	🥇	0.915	✗		**0.961**	✓	🥇
predict trans. conductors		Min ↓	0.069	0.065	0.062	0.055	0.294	✗		0.183	✗		0.064	✓	🥉	**0.062**	✓	🥈
english text normalization		Max ↑	0.990	0.990	0.991	0.997	0.0	✗		NR	✗		0.834	✗		**0.990**	✓	🥉
may 2022 tab playground		Max ↑	0.972	0.998	0.998	0.998	0.711	✗		0.882	✗		0.987	✓		**0.992**	✓	
random acts of pizza		Max ↑	0.599	0.692	0.724	0.979	0.520	✗		0.591	✗		**0.655**	✓		**0.643**	✓	
spooky author identification		Min ↓	0.418	0.293	0.269	0.165	0.992	✗		0.582	✗		0.320	✓		**0.532**	✗	
jigsaw toxic comments		Max ↑	0.980	0.986	0.986	0.987	0.570	✗		0.970	✗		**0.984**	✓		0.874	✗	
russian text normalization		Max ↑	0.975	0.975	0.982	0.990	0.486	✗		0.486	✗		**0.920**	✗		0.000	✗	
NYC taxi fare prediction		Min ↓	3.597	2.923	2.881	2.337	1.2e13	✗		355.8	✗		10790	✗		**6.542**	✗	