	resource intensity	scalability	automation potential	real-world applicability	numerical reducibility
traditional benchmarks	<b>✓</b>	<b>✓</b>	<b>✓</b>	×	<b>✓</b>
challenges and competition	_	_	_	<b>✓</b>	_
red teaming and capability discovery	X	X	_	<b>✓</b>	X
real-world deployment studies	X	X	X	<b>✓</b>	X
ablation studies and systematic (stress) testing	X	<b>✓</b>	<b>✓</b>	_	<b>✓</b>