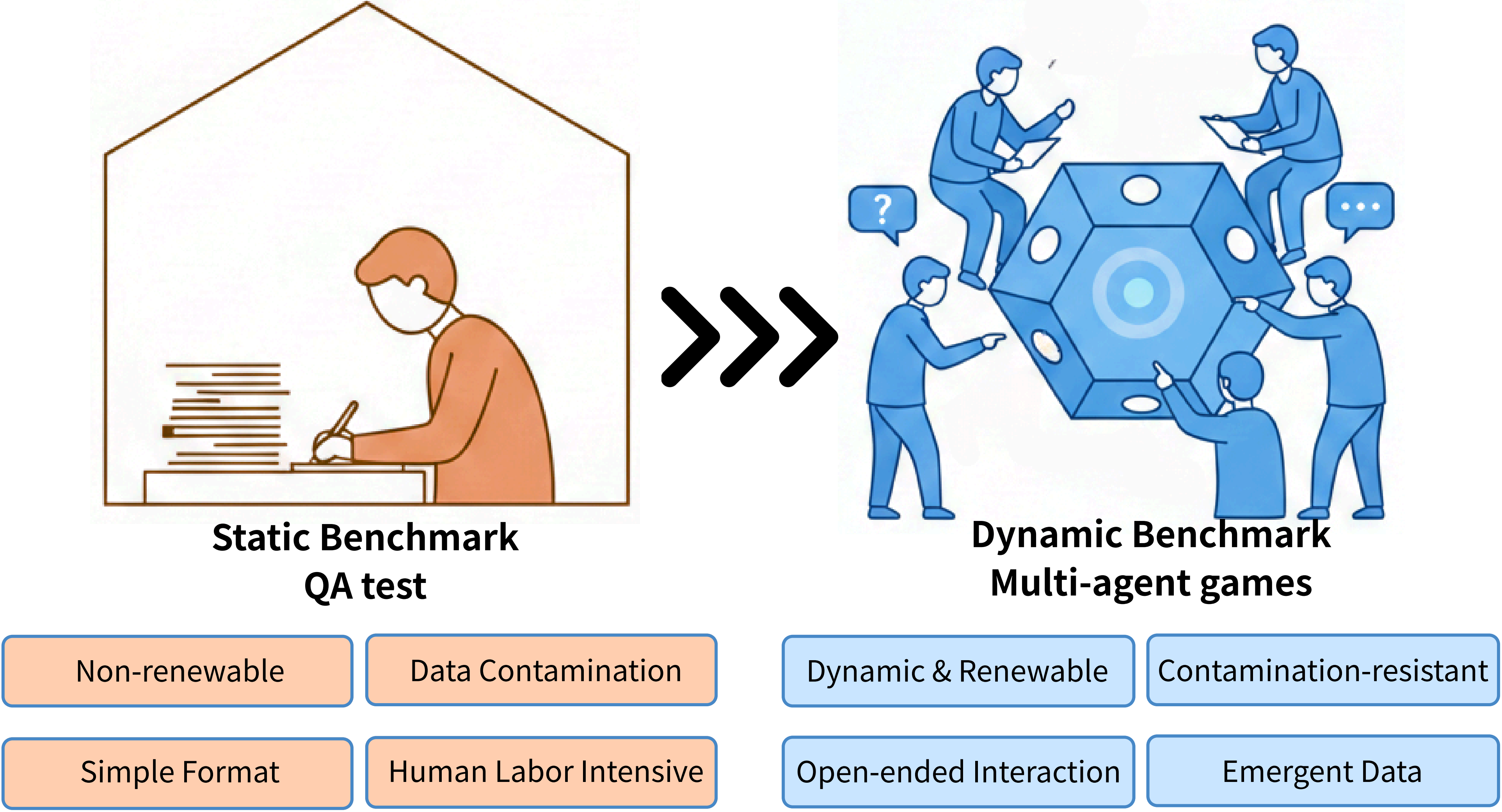
Static Benchmark
QA test
Dynamic Benchmark
Multi-agent games
Non-renewable
Data Contamination
Simple Format
Human Labor Intensive
Dynamic & Renewable
Contamination-resistant
Open-ended Interaction
Emergent Data