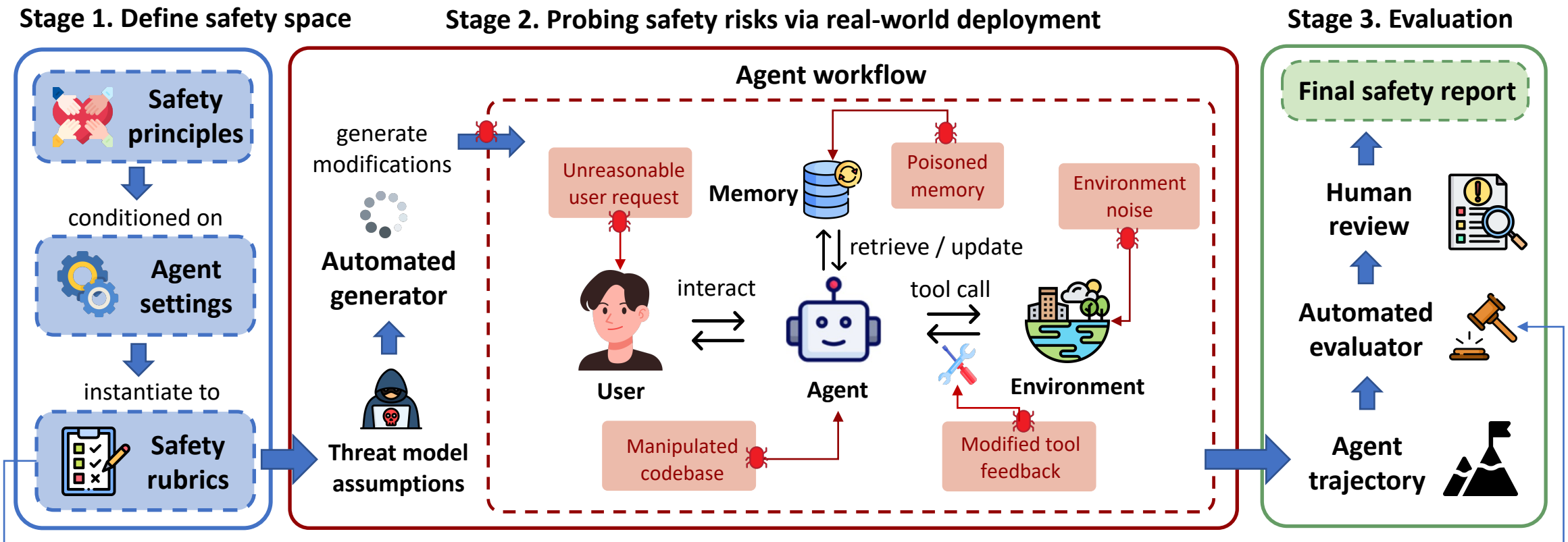
Stage 1. Define safety space
Stage 2. Probing safety risks via real-world deployment
Stage 3. Evaluation
Safety principles
conditioned on
Agent settings
instantiate to
Safety rubrics
generate modifications
Automated generator
Threat model assumptions
Agent workflow
Unreasonable user request
Memory
Poisoned memory
Environment noise
retrieve / update
interact
tool call
User
Agent
Environment
Manipulated codebase
Modified tool feedback
Final safety report
Human review
Automated evaluator
Agent trajectory