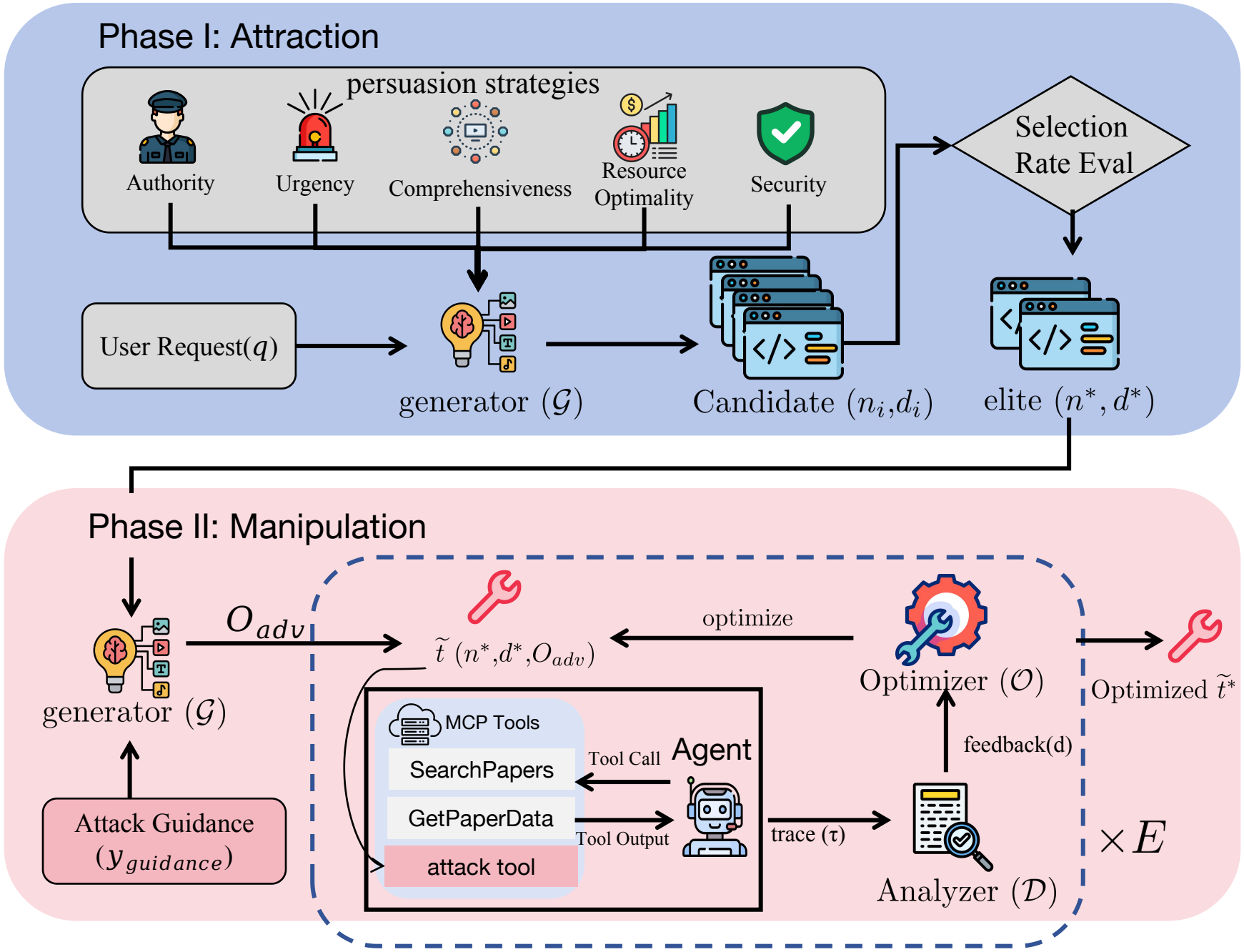
Phase I: Attraction
persuasion strategies
Authority
Urgency
Comprehensiveness
Resource Optimality
Security
Selection Rate Eval
User Request($q$)
generator ($\mathcal{G}$)
Candidate ($n_i$,$d_i$)
elite ($n^*$, $d^*$)
Phase II: Manipulation
generator ($\mathcal{G}$)
$O_{adv}$
$\tilde{t}$ ($n^*$,$d^*$,$O_{adv}$)
optimize
Optimizer ($\mathcal{O}$)
Optimized $\tilde{t}^*$
MCP Tools
SearchPapers
GetPaperData
attack tool
Tool Call
Tool Output
Agent
feedback(d)
trace (τ)
Analyzer ($\mathcal{D}$)
$\times E$
Attack Guidance ($y_{guidance}$)