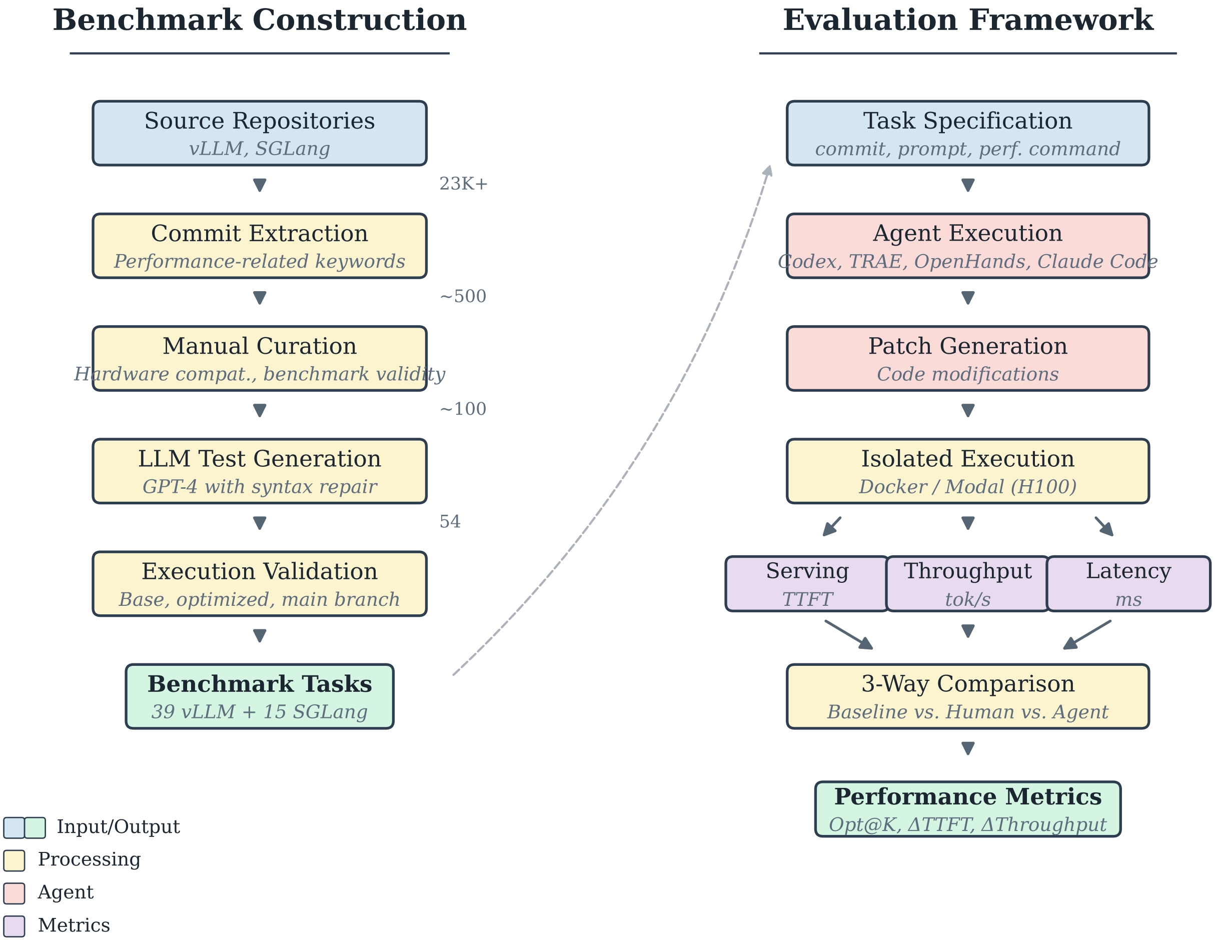
Benchmark Construction
Source Repositories
vLLM, SGLang
23K+
Commit Extraction
Performance-related keywords
~500
Manual Curation
Hardware compat., benchmark validity
~100
LLM Test Generation
GPT-4 with syntax repair
54
Execution Validation
Base, optimized, main branch
Benchmark Tasks
39 vLLM + 15 SGLang
Evaluation Framework
Task Specification
commit, prompt, perf. command
Agent Execution
Codex, TRAE, OpenHands, Claude Code
Patch Generation
Code modifications
Isolated Execution
Docker / Modal (H100)
Serving
TTFT
Throughput
tok/s
Latency
ms
3-Way Comparison
Baseline vs. Human vs. Agent
Performance Metrics
Opt@K, ΔTTFT, ΔThroughput
Input/Output
Processing
Agent
Metrics