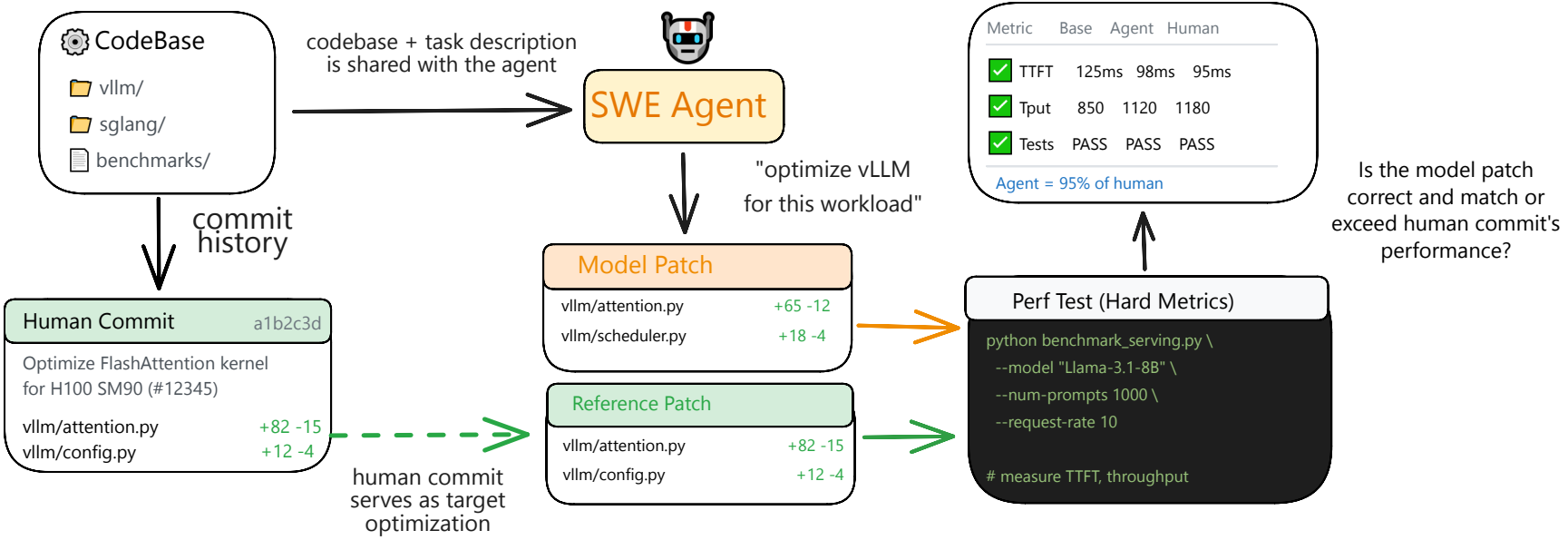
CodeBase
vllm/
sglang/
benchmarks/
codebase + task description is shared with the agent
SWE Agent
commit history
"optimize vLLM for this workload"
Metric Base Agent Human
TTFT 125ms 98ms 95ms
Tput 850 1120 1180
Tests PASS PASS PASS
Agent = 95% of human
Is the model patch correct and match or exceed human commit's performance?
Human Commit a1b2c3d
Optimize FlashAttention kernel for H100 SM90 (#12345)
vllm/attention.py +82 -15
vllm/config.py +12 -4
Model Patch
vllm/attention.py +65 -12
vllm/scheduler.py +18 -4
Perf Test (Hard Metrics)
python benchmark_serving.py \
--model "Llama-3.1-8B" \
--num-prompts 1000 \
--request-rate 10
measure TTFT, throughput
human commit serves as target optimization
Reference Patch
vllm/attention.py +82 -15
vllm/config.py +12 -4