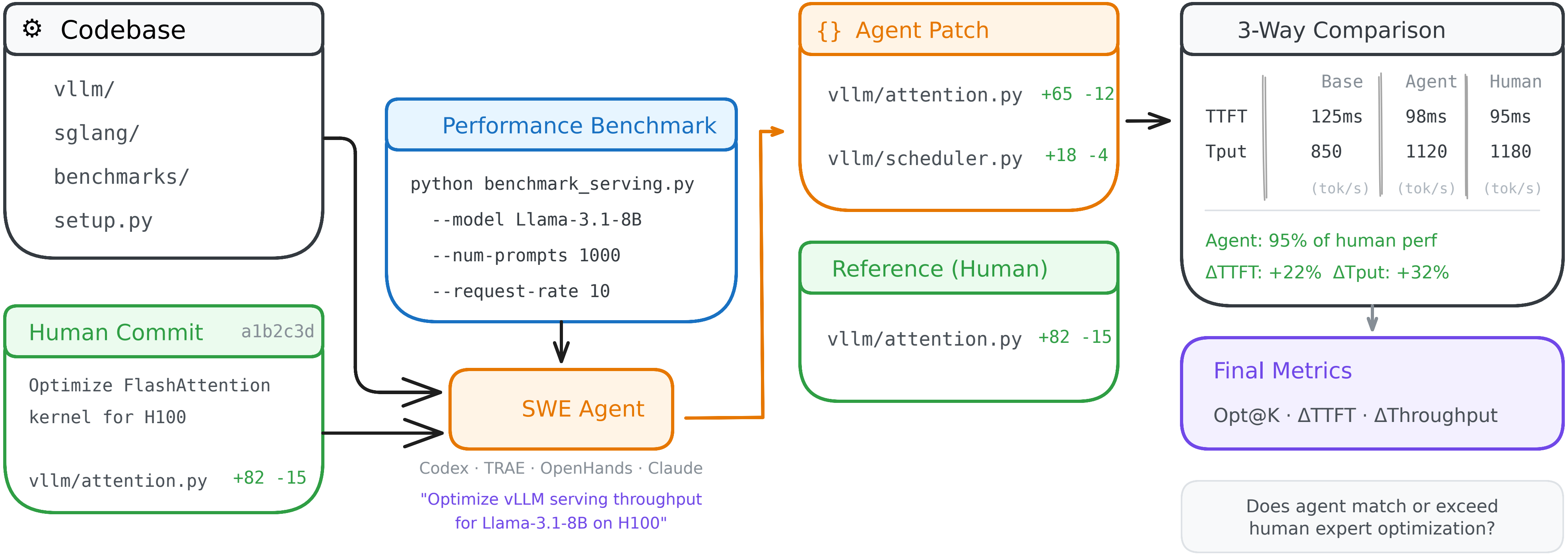

⚙ Codebase

```
vllm/
sglang/
benchmarks/
setup.py
```

Human Commit a1b2c3d

```
Optimize FlashAttention
kernel for H100

vllm/attention.py  +82 -15
```

Performance Benchmark

```
python benchmark_serving.py
  --model Llama-3.1-8B
  --num-prompts 1000
  --request-rate 10
```

SWE Agent

Codex · TRAE · OpenHands · Claude

"Optimize vLLM serving throughput for Llama-3.1-8B on H100"

{} Agent Patch

```
vllm/attention.py   +65 -12
vllm/scheduler.py   +18 -4
```

Reference (Human)

```
vllm/attention.py  +82 -15
```

3-Way Comparison

	Base	Agent	Human
TTFT	125ms	98ms	95ms
Tput	850	1120	1180
	(tok/s)	(tok/s)	(tok/s)

Agent: 95% of human perf

ΔTTFT: +22% ΔTput: +32%

Final Metrics

Opt@K · ΔTTFT · ΔThroughput

Does agent match or exceed human expert optimization?