

AgentPProf: Semantic Profiler for Long Horizon AI Agents

Yusheng Zheng · Chaokun Chang · Yu Mao · Tianyuan Wu · Yuxi Huang · Tao Ma · Wenan Mao · Shuyi Cheng · Andi Quinn · Wei Wang

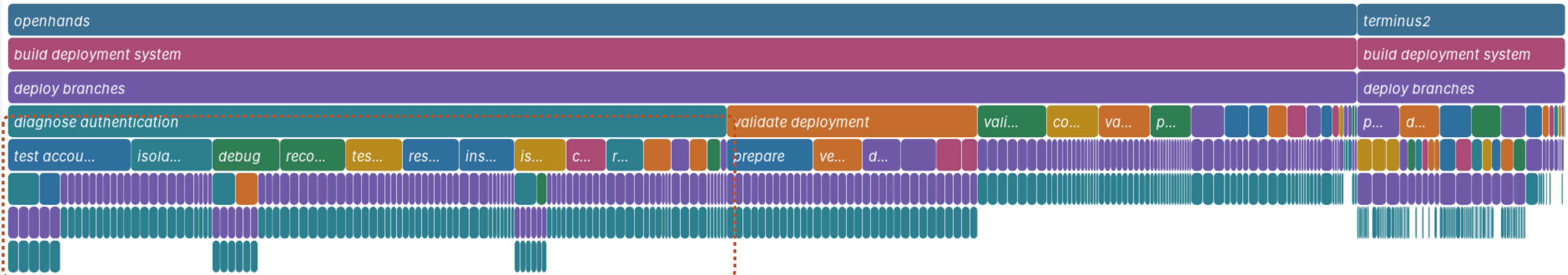
github.com/eunomia-bpf/agentsight

AgenticOS 2026, Prague

1 PROBLEM

Agents need profiling, not only debugging

Agents run for hours or days. Teams ask questions across many runs: which tasks consume the most tokens, where failures concentrate. Tracing shows one run; profiling aggregates cost and attributes it to what is responsible.



Three failed Git-deployment runs, folded (width = tokens): *diagnose authentication* takes **46%** of tokens but 21% of operations; agents retry SSH variants.

But no trajectory contains the name *diagnose authentication*.

2 CHALLENGE

No call stack, no stable names

Task intent, not code paths. Prompts are natural language: the three runs say *diagnose authentication* in different words.

No hierarchy. A few prompts may be one task or several; nothing marks the boundaries.

The profiling methodology still transfers. **Needed, without a call stack:**

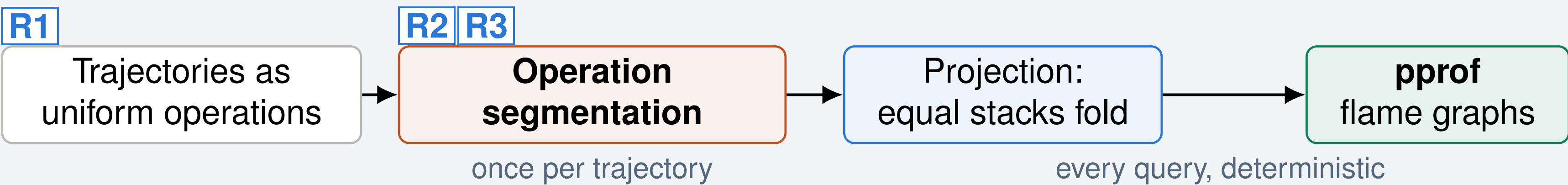
R1 system effects linked to task intent

R2 stable, repeatable names

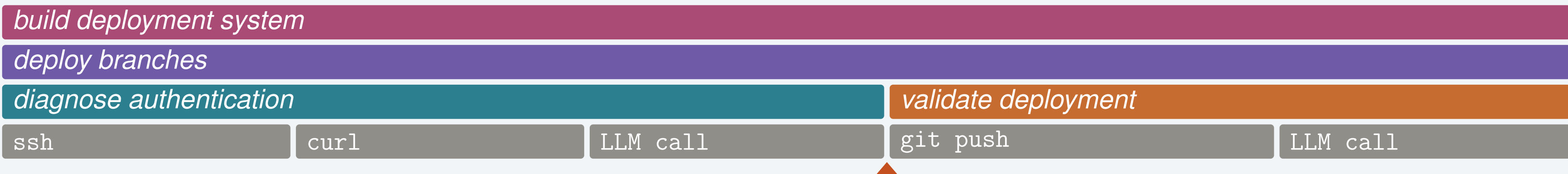
R3 a hierarchy built from data

3 AGENTPPROF

Every activity is an operation; tasks form its stack



Recursive operation segmentation. A task occupies a contiguous span, so only the **transitions** need to be found: cut, name each part with a short verb phrase, recurse.

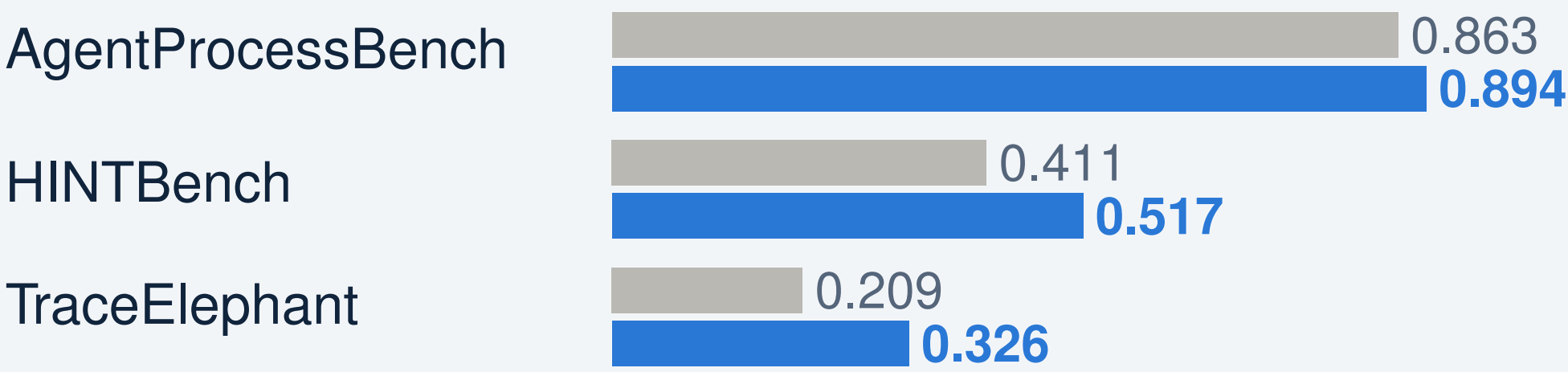


Schematic (names from Figure 1c): equal paths fold across runs; LLM and tool calls stay as leaf evidence; ▲ marks a transition.

4 EVALUATION

From profile to fix. An analyst agent given only the pprof output locates a ToolSandbox bug; the one-line fix cuts agent tokens by **19%** without degrading task quality.

Fault localization, MAP (0–1, how high the true faulty step ranks): the profile adds **4–56%**.



gray: benchmark's own judge · blue: + AgentPProf profile

Segments match human stages: AI agent segmentation reaches B³ F1 **0.764**, vs. 0.663 for the strongest baseline.

Cost: segmentation once, 37 min for 405 trajectories; a 27,765-operation profile then builds in **1.16 s**.

A semantic operation stack model adapts profiling to agent trajectories; AgentPProf aggregates them into pprof-compatible profiles.