

Multi-Agent Collaboration Benchmark for Software Engineer



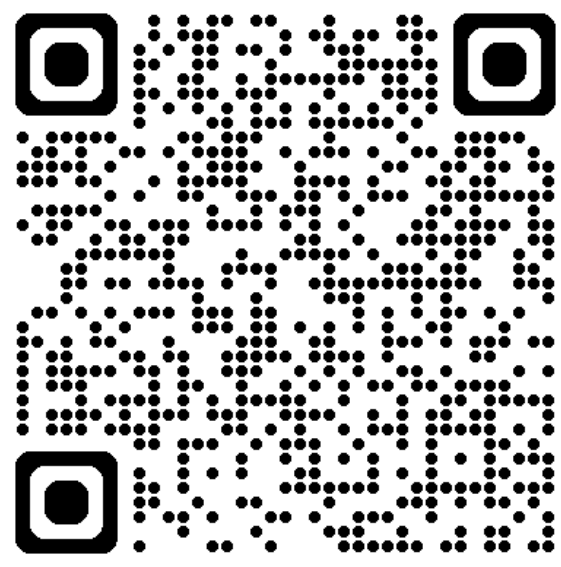
Academic Supervisor
Nguyen Minh Long



Industry Supervisor
Vu Bui
Uyen Pham

TheOrchestrator

Tran Phan Anh Khoa s4136776
Nguyen Hoang Long s4131459
Doan Dac Nguyen s4131473
Cao Ngoc Phuong Uyen s3919659



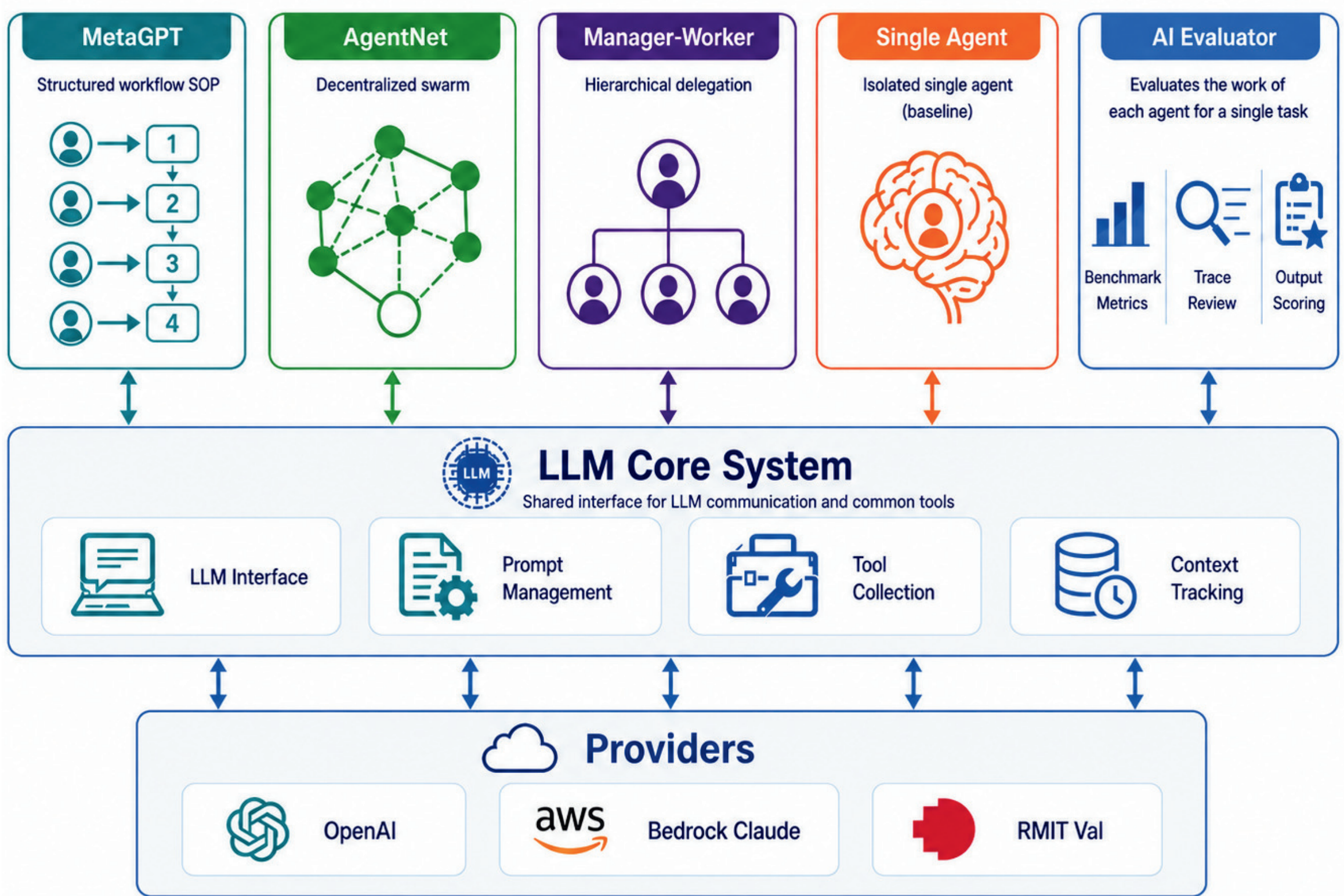
Background

- Modern continuous integration pipelines demand frequent software releases, rendering manual testing obsolete.
- Current single-agent AI systems become processing bottlenecks when handling complex Software Quality Assurance (SQA) workflows.
- While multi-agent systems solve this through a specialized division of labor, evaluating their non-deterministic group behavior remains a critical challenge.
- This project transitions from a standalone web platform to a reusable backend engine designed for seamless integration into Katalon's existing True Platform ecosystem.

Objectives

- Strategy Engine:** Execute configurable multi-agent collaboration processes that support remote external agents
- Evaluation Engine:** Measure task outcomes, collaboration efficiency, token consumption, and execution time
- Empirical Comparisons:** Contrast four distinct execution paradigms: the structured SOP workflows of MetaGPT, the hierarchical delegation of Manager-Worker, the decentralized adaptability of AgentNet, and isolated Single-Agent systems.

Backend Architecture



Methodology

1. Strategy Stabilization

- Implement strategy modules
- Optimize prompts using inside the strategies.
- Refine execution to avoid unexpected errors crashing mid-way.

2. Metric Design

- Define metrics for task performance, collaboration behaviour, efficiency, and reliability
- Implement evaluator module

3. Data Collection & Normalization

- Refine each strategy to capture necessary execution traces.
- Normalize the traces to the evaluator input

4. Benchmark Test Suite

- Design benchmark tasks covering different task types, complexities, and collaboration requirements

5. Controlled Experiments

- Build a controlled environment to be able to run test cases repeatedly for each strategy.
- Run the test suites in this controlled environment.

6. Analysis & Visualization

- Collect metrics and save them to database
- Build a dashboard to visualize the metrics value for better analysis

Result

