AgentQuest: A Modular Benchmark Framework to Measure Progress and Improve LLM Agents

Luca Gioacchini, Giuseppe Siracusano, Davide Sanvito, Kiril Gashteovski, David Friede, Roberto Bifulco, Carolin Lawrence

Benchmarking Generative AI Agents

- **Generative AI Agents**: software systems leveraging LLMs to perform complex multi-steps tasks
  - Key concepts: *Environment, State, Observation, Action*
- Benchmarking is essential for evaluation and guiding improvements
- **Issues in current benchmarks:**
  - Limited coverage of different types of benchmarks
  - Existing metrics focus on final success:
    - ***Success Rate (SR)*** and ***Time to Success (Steps)***
  - Coupling of benchmarks and agent architectures

❌Closed-box vs open-ended tasks with tools usage
❌Limited insights into *intermediate* success/failure
❌Limited re-usability and extensibility

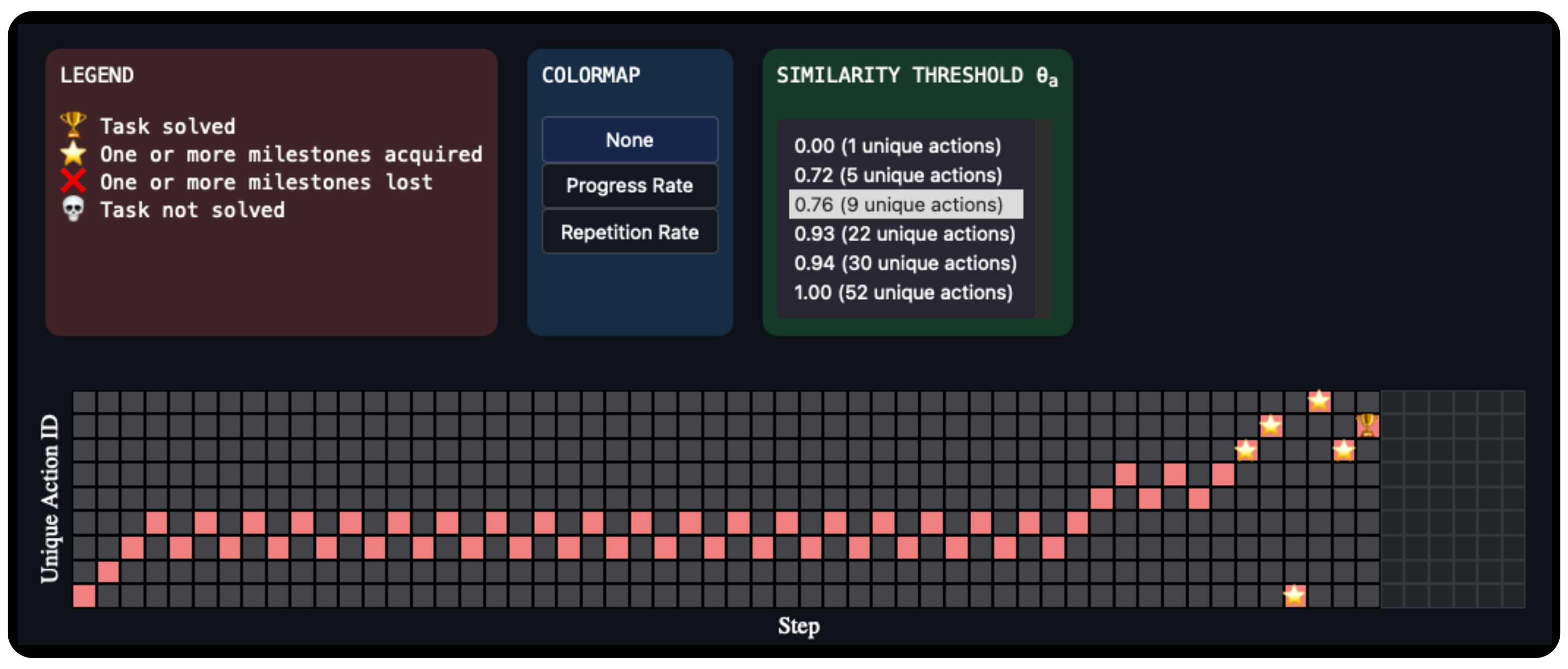
AgentQuest

- Moudular framework to support multiple diverse benchmarks and agent architectures
- Single unified Python interface
- Two new cross-benchmarks metrics:
  - ***Progress Rate (PR)*** and ***Repetition Rate (RR)***
- Enables the **investigation of agent behaviour** to support the improvement of its architecture
- (Initial) support for 4 benchmarks:
  - Evaluation of an agent based on LangChain Chat Model powered by GPT-4
  - ***Improved results** after modifying the agent architecture **based on AgentQuest's insights**
- Much more in the paper and in the GitHub repository!
  - Additional agent architectures and open-ended benchmark

✅Agent / Environment decoupling
✅Increased progress observability
✅Simplified interoperability across benchmarks and agent architectures and improved extensibility

	Existing Metrics		AgentQuest	
	SR	Steps	PR_{60}	RR_{60}
Mastermind	0.47	41.87	0.62	0.32
LTP	0.20	52.00	0.46	0.81
ALFWorld	0.86	21.00	0.74	0.06
Sudoku	0.00	59.67	0.08	0.22
Mastermind*	0.60 +13 %pts	39.73	0.73 +11 %pts	0.00
ALFWorld*	0.93 +7 %pts	25.86	0.80† +6 %pts	0.07†

†Metrics with extended runtime up to 120 steps, i.e. PR_{120} and RR_{120}.



LEGEND

🏆 Task solved
⭐ One or more milestones acquired
❌ One or more milestones lost
💀 Task not solved

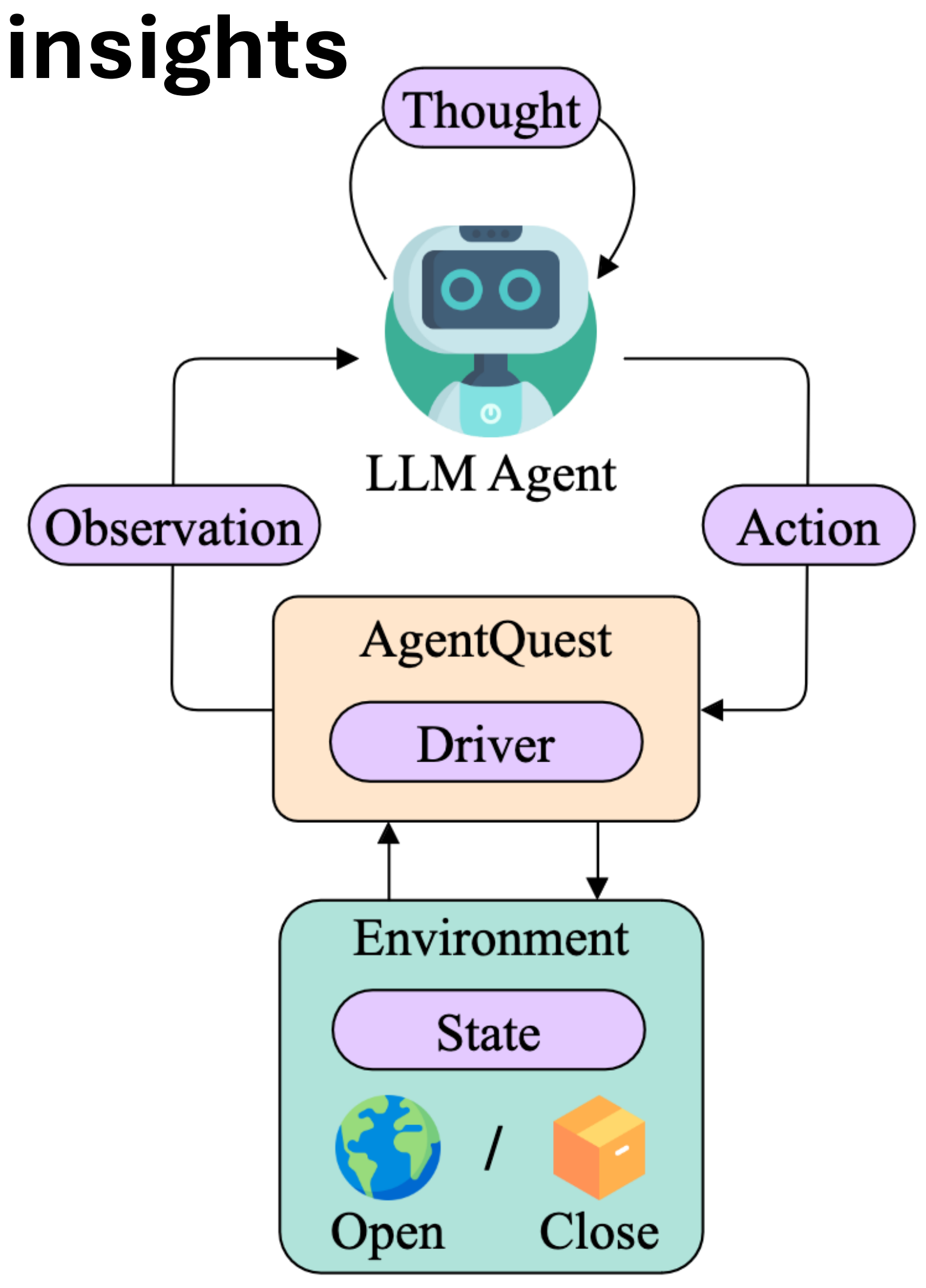
COLORMAP

None
Progress Rate
Repetition Rate

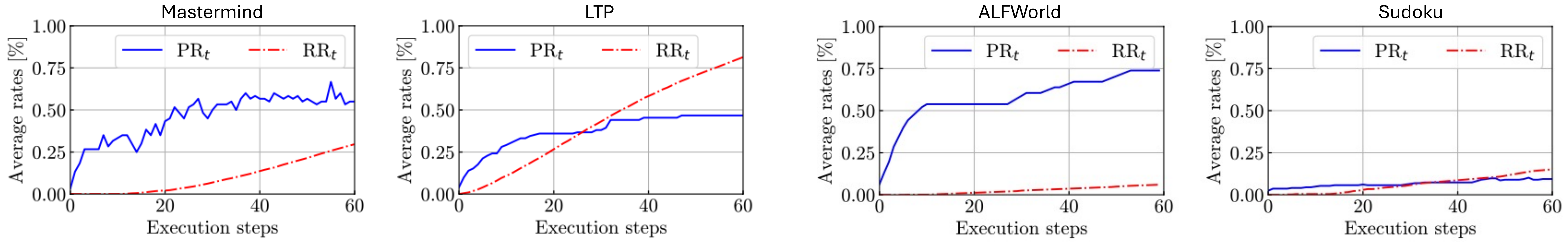
SIMILARITY THRESHOLD θ_a

0.00 (1 unique actions)
0.72 (5 unique actions)
0.76 (9 unique actions)
0.93 (22 unique actions)
0.94 (30 unique actions)
1.00 (52 unique actions)

The **AgentQuest GUI** provides a convenient visualization to track the progress over time and perform a step-by-step investigation of the actions performed by the agent.



AgentQuest high-level architecture.



AgentQuest enables to monitor progress and action repetition to get relevant insights to improve agent architecture: e.g. adding a memory component (for Mastermind) or extending the benchmark runtime (for ALFWorld).

AgentQuest is open-source!

- 🌐 https://github.com/nec-research/agentquest
- 📖 Documentation for existing benchmarks
- 💡 Examples based on OpenAI and LangChain APIs
- 🚀 HOWTOs for new benchmarks, drivers and metrics

VIDEO

CODE

NEC

Politecnico di Torino

1859

This project has received funding from the European Union's Horizon Europe research and innovation programme (SNS-JU) under the Grant Agreement No 101139285 ("NATWORK").

NAACL 2024