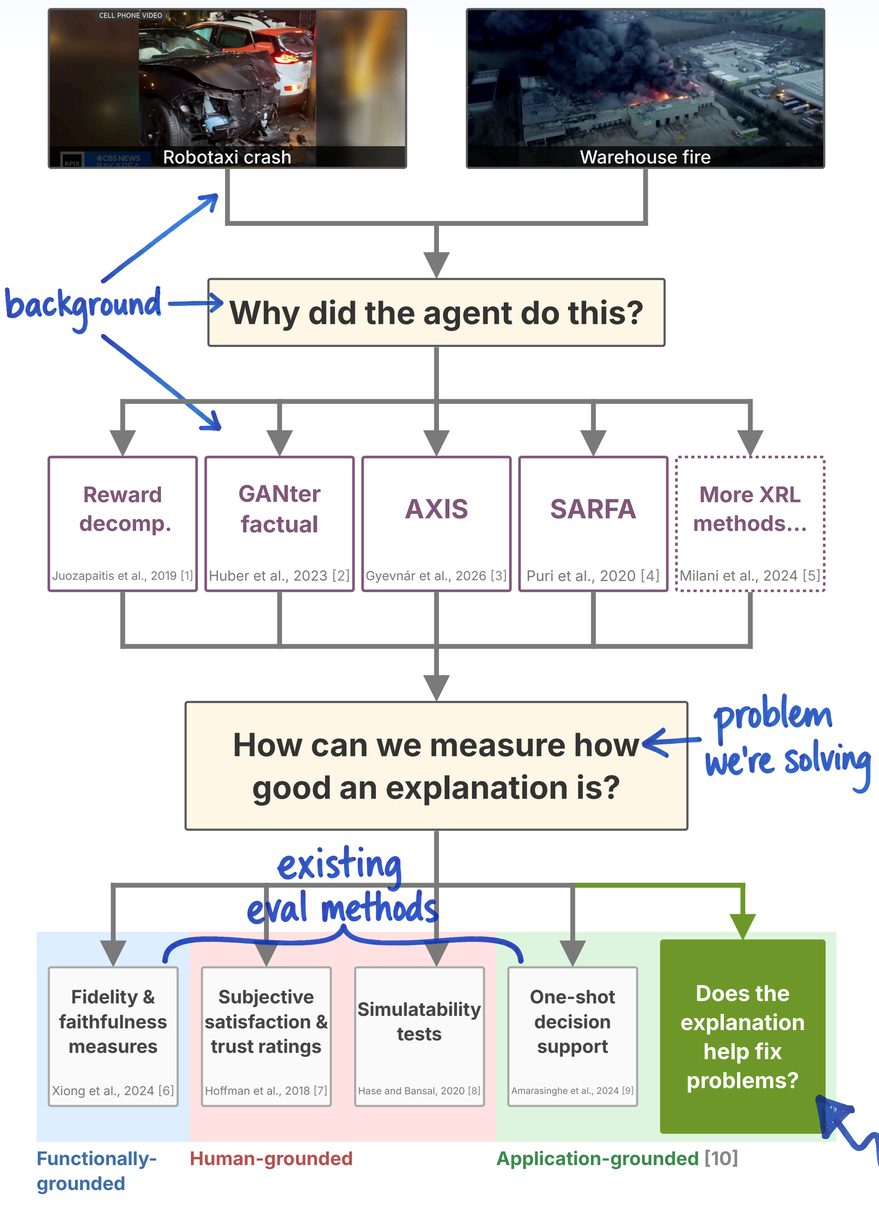
EvalXRL: Evaluating XRL by Fixing Broken Agents

Center for Human-Compatible Artificial Intelligence

Ram Rachum[1,2] Yotam Amitai[3] Bálint Gyevnár[4] Reuth Mirsky[2] Cameron Allen[1]

[1]UC Berkeley [2]Tufts University [3]Independent Researcher [4]Carnegie Mellon University

GOLD Lab



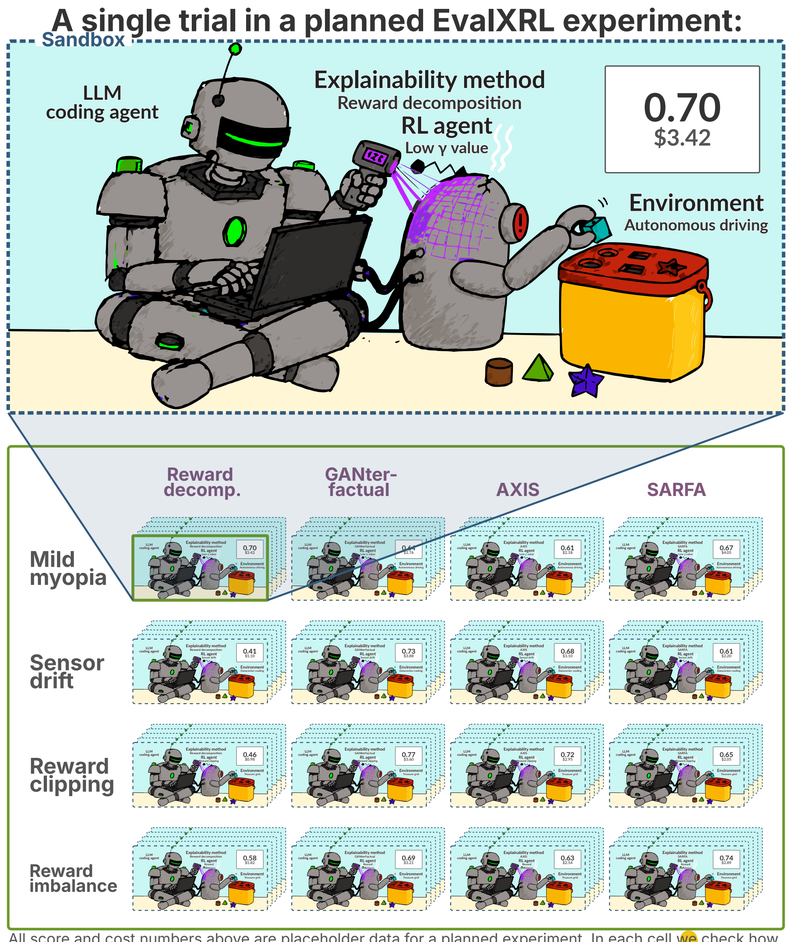
We present a new evaluation paradigm for RL explainability methods. We plan EvalXRL as an application-grounded benchmark [10]: we intentionally break an RL agent in a secret way, then ask a separate LLM coding agent to diagnose what is wrong with the RL agent, and then fix it.

We equip the coding agent with an explanation method, that may be used to glean useful information. We sample (1) the post-repair performance of the RL agent and (2) the token spend, and report both as performance attributes of the explainability method.

Paper: r.rachum.com/toward-evalxrl

Live demo: fixarena.org

A single trial in a planned EvalXRL experiment:

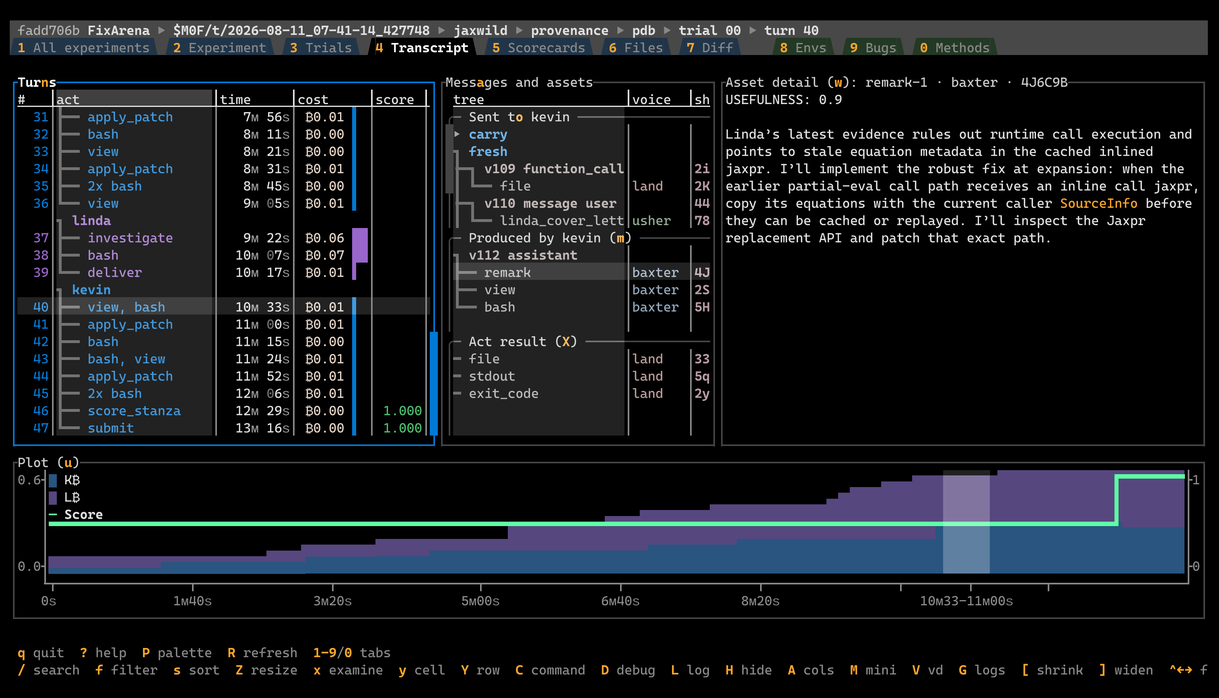


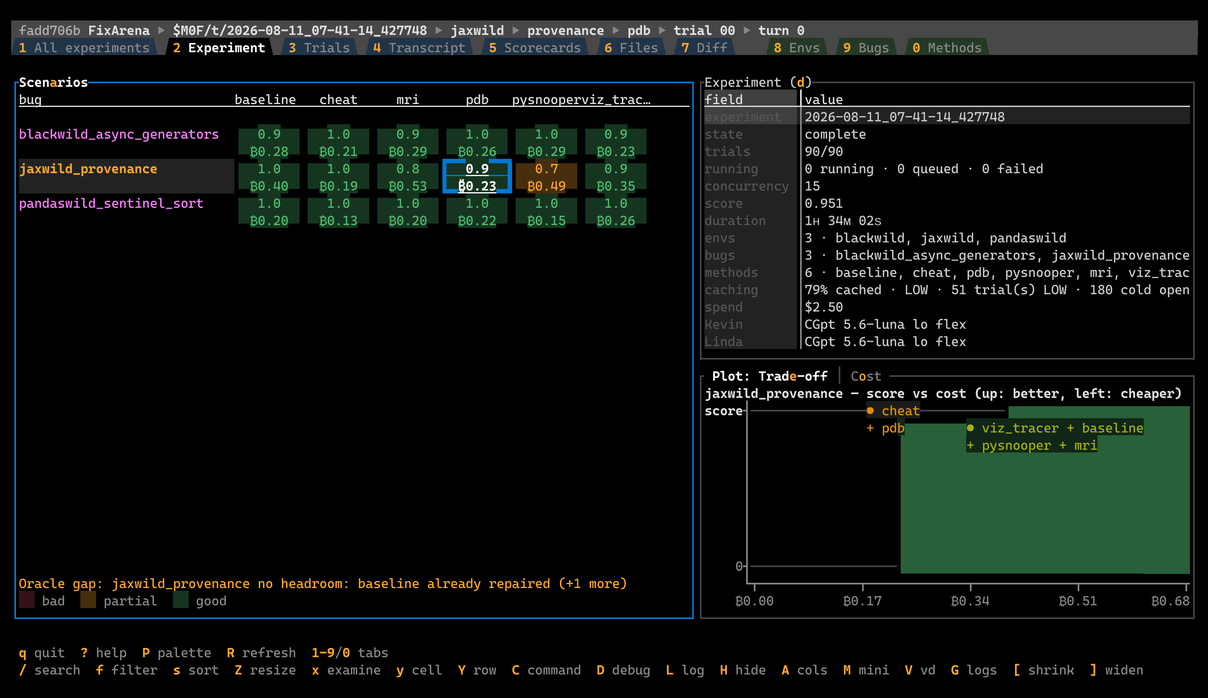
All score and cost numbers above are placeholder data for a planned experiment. In each cell we check how much one particular explainability method has helped the LLM coding agent fix one particular RL problem. We aggregate these scores and report them as performance attributes of the explainability method.

FixArena: Framework for Measuring LLM Coding Assistance

FixArena is our work-in-progress infrastructure. When it's ready, we plan to use it for running EvalXRL experiments, and then apply the same approach to Mechanistic Interpretability of LLMs.

It orchestrates hundreds of Docker sandboxes, iterating over different misalignments and different methods. It measures how successfully the LLM coder fixes the RL agent, and then aggregates that data to plots comparing the success of different explanation methods.





[1] Z. Juozapaitis et al. Explainable RL via Reward Decomposition. IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019. [2] T. Huber et al. GANterfactual-RL: Understanding Reinforcement Learning Agents' Strategies through Visual Counterfactual Explanations. AAMAS 2023. [3] B. Gyevnár et al. Integrating Counterfactual Simulations with Language Models for Explaining Multi-Agent Behaviour. AAMAS 2026. [4] N. Puri et al. Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution. ICLR 2020. [5] S. Milani et al. Explainable Reinforcement Learning: A Survey and Comparative Review. ACM Computing Surveys 2024. [6] Y. Xiong et al. XRL-Bench: A Benchmark for Evaluating and Comparing Explainable RL Techniques. KDD 2024. [7] R. Hoffman et al. Metrics for Explainable AI: Challenges and Prospects. arXiv 2018. [8] P. Hase and M. Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? ACL 2020. [9] K. Amarasinghe et al. On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods. AAAI 2024. [10] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable ML. arXiv 2017.