MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science

Ran Xu[1], Yuchen Zhuang[2], Yishan Zhong[2], Yue Yu[2], Zifeng Wang[3], Xiangru Tang[4], Hang Wu[2], May D. Wang[2], Peifeng Ruan[5], Donghan Yang[5], Tao Wang[5], Guanghua Xiao[5], Xin Liu[6], Carl Yang[1], Yang Xie[5†], Wenqi Shi[5†]

[1]Emory University [2]Georgia Institute of Technology [3]University of Illinois Urbana-Champaign [4]Yale University [5]UT Southwestern Medical Center [6]University of Washington

Overview

◇ We introduce `MedAgentGym`, **a scalable and interactive training environment** designed to enhance coding-based biomedical reasoning capabilities in LLM agents.
◇ `MedAgentGym` comprises **72,413** task instances across **129** categories derived from **12** authentic real-world biomedical scenarios.
◇ Tasks are encapsulated within **executable sandbox environments**, each featuring detailed task specifications, interactive feedback mechanisms, verifiable ground truth annotations, and scalable training trajectory generation.
◇ Extensive benchmarking of **29 LLMs** reveals substantial performance disparities in biomedical data science between commercial and open-source LLMs.
◇ Leveraging efficient multi-threaded and multi-turn trajectory sampling in `MedAgentGym`, `Med-Copilot` achieves performance gains of **+43.02%** and **+45.28%** from offline and online reinforcement learning, respectively, demonstrating `MedAgentGym` as an effective training ground while establishing itself as a cost-effective, privacy-preserving alternative competitive with proprietary LLMs.

Framework

`MedAgentGym` provides an interactive coding environment for LLM agents across diverse biomedical data science tasks.

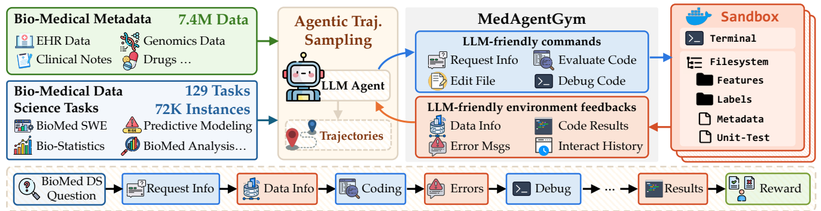


Figure: Overview of `MedAgentGym`.

Key Infrastructure:

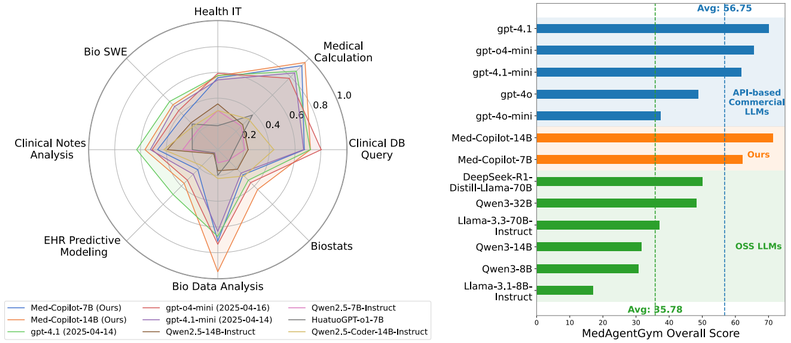
◇ **Isolated Docker sandboxes** with pre-installed biomedical dependencies, ensuring environmental integrity and medical data security;

◇ **Interactive feedback**: robust JSON parsing and debugging with error grounding, translating compile-time and runtime errors into unified natural language format;

◇ **Efficient trajectory collection** via `Ray` and `Joblib` multi-threading backends, supporting parallel execution and sequential sampling;

◇ **Plug-and-play** modular architecture readily supporting new biomedical coding task integration through configuration files.

Benchmarks: LLMs as Biomedical Data Science Agents



(a) Task-specific biomedical coding capabilities (b) Overall score leaderboard

Figure: Overview of (a) task-specific and (b) overall leaderboard evaluation in `MedAgentGym`.

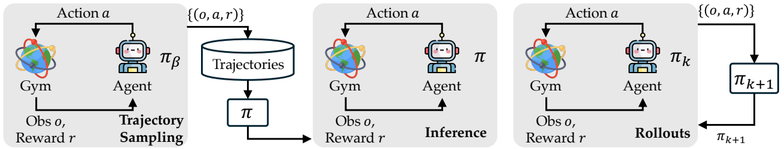
Key Findings:

◇ **Significant performance gap** (>20%) between commercial API-based and OSS LLMs, highlighting the *critical need* to develop lightweight OSS models.

◇ LLMs perform better in structured tasks (*e.g.*, database queries) compared to open-ended tasks requiring advanced coding (*e.g.*, ML prediction).

◇ Both coding (♀) and medical (♂) LLMs deliver **limited improvement or even decline** over base models, revealing that coding-based biomedical reasoning is a *unique capability* not captured by either specialization alone.

Training LLMs for Coding-Centric Biomedical Reasoning



(a) Offline Reinforcement Learning w/ MedAgentGym (b) Online RL w/ MedAgentGym

Figure: Comparison of (a) offline and (b) online RL paradigms within `MedAgentGym`.

Experiments: Training Med-Copilot with Agentic RL

Table: `Med-Copilot` performance on `MedAgentGym` fine-tuned with sampled trajectories.

Datasets (→) / Base (↓) / Metrics (→)	MIMIC SR	eICU SR	TREQS SR	MedCalc SR	MedAg. SR	BioC. SR	BioDS. SR	EHRS. Acc	Avg. Score	Δ
`Qwen2.5-7B-Inst`	13.08	15.57	12.76	25.91	30.36	21.79	10.20	5.42	16.89	–
+SFT	57.83	61.48	72.66	89.06	50.85	28.33	55.10	15.62	53.87	(+36.98)
+DPO	64.13	66.91	72.02	90.06	52.54	34.62	69.39	29.55	59.90	(+43.02)
+PPO	66.10	67.25	**73.88**	74.52	51.33	32.71	65.47	32.40	57.96	(+41.07)
+GRPO	**68.21**	**68.73**	70.50	**92.33**	**55.87**	**37.40**	**71.11**	**33.18**	**62.17**	**(+45.28)**
`Qwen2.5-14B-Inst`	17.21	14.07	16.43	27.40	35.59	29.49	16.33	4.45	20.12	–
+SFT	61.45	62.46	76.38	94.36	52.54	39.80	89.80	34.58	63.92	(+43.80)
+DPO	64.54	63.52	76.08	92.45	54.32	43.56	92.96	43.56	66.37	(+46.25)
+PPO	67.55	68.53	**78.32**	94.86	53.22	45.88	91.33	56.79	69.56	(+49.44)
+GRPO	**68.78**	**69.34**	76.84	**95.81**	**57.41**	**49.32**	**94.78**	**59.05**	**71.42**	**(+51.30)**

Experiments: Scaling, Ablation, and Error Analysis

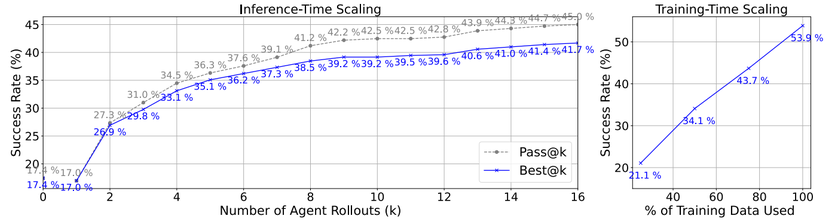
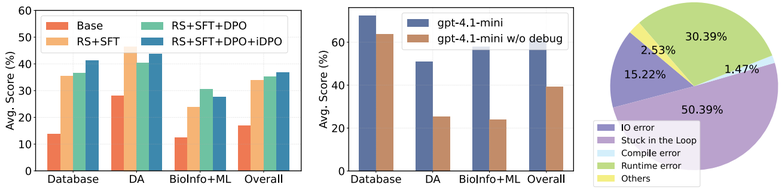


Figure: Scalable improvements of LLM agents in `MedAgentGym`. For inference-time scaling, we employ $T = 0$ for the initial rollout and $T = 0.6$ for the rest. For train-time scaling, we set $T = 0$.



(a) Self-Improvement (b) Effect of Debug (c) Error Types