Co-Evolution of Scientific Tools and Agents

Yang Zhang, Peking University

I. Motivation

As capabilities of Large Language Models (LLMs) continue to advance, a fundamental question emerges in AI for Science: *how can we continuously translate these advancing capabilities into deeper, more scalable engagement in scientific discovery?*

Tool-usage remains a key perspective to answering this question. Scientific agents have been apt at using tools, whether as necessary steps of some tasks, e.g. running verification in physical design, or as complements to their own short-comings, e.g. running information retrievers or solvers. However, to advance scientific agents into highly-specialized domains or reasoning-intensive tasks such as discovery, the following challenges persist:

(1) **Tool Creation.** Many current tool creation pipelines primarily focus on transferring and adapting existing programs [1]–[3], or generating tools from structured knowledge (e.g. organized open-source code bases [4], extracted symbolic representations [5]), leading to dependency on the availability of prior solutions or well-organized knowledge bases, which limits effectiveness in lesser-explored or novel scientific domains. In contrast, a substantial portion of scientific knowledge resides in unstructured sources, such as reference books and web corpora, offering rich yet largely untapped opportunities for the evolution of scientific agents.

(2) **Tool Evolution.** The structuring and quality-control of tool sets are key to elevating and generalizing scientific agents' performances. Seminal works like TroVE [6] and KTCE [5] develop a self-evolving tool set by focusing on reusability, diversity and retrieval-efficiency. However, these frameworks treat self-evolution as a refinement mechanism, without explicitly studying how to modify tool sets in response to newly acquired domain knowledge or feedback from deployment, which is a necessary task for comprehensive scientific agents, and for transferable rather than ad-hoc tool sets.

(3) **Agent Evolution.** Although LLM agents have made great progress in tool-usage through various instruction-tuning [7] and RL-based methods [8], [9], their performance remain limited when it comes to the cost-awareness of tool-usage [10], full-stack research [11] and safety of scientific tool-usage [12]. These limitations dictate training approaches and benchmarks tailored to co-evolution with tools.

II. Roadmap

In response to these challenges, I propose to study the **Co-Evolution of Scientific Tools** (LLMs' abilities to create and evolve tools) **and Agents** (LLMs' abilities to utilize tools for practical deployment in scientific tasks). Advanced LLM-governed tool sets will significantly improve LLMs' competence as scientific agents, and elevated agentic capabilities will allow for more diverse, high-quality tools. Concretely, I propose the following directions:

(1) *Evaluation.* Current tool-related benchmarks focus on various aspects of tool-utilization [13], [14], but few explicitly evaluate agents' abilities to solve scientific problems by creating tools, or to use scientific tools under domain-specific constraints and goals. However, these scenarios are prevalent, especially in scientific domains that are underrepresented in LLM pretraining and lack existing callable tools. I propose to evaluate the ability to create, govern and utilize tools. Such abilities should be depicted by verifiable metrics such as tool-quality and tool-transferability at creation, and cost-efficiency and robustness at utilization.

(2) *Tool Generation from Unstructured Material.* To address challenge (1) and (2) in Section I, a tool-creation and tool-governing agentic framework that takes in unstructured knowledge (textbooks, tutorials, web information, etc.) and produces structured tool set is needed. Moreover, it should incorporate frequent knowledge updates / tool feedback as in authentic scientific scenarios. Such a framework could potentially perform well in zero-shot, highly-specified scientific settings.

(3) *RLVR.* Given challenges in Section I, training agents that design and utilize tools is a non-trivial task. I plan to utilize RL-based approaches with verifiable signals as proposed in direction (1) (*Evaluation*). With stage-wise training and fast verifications (e.g. surrogate models to estimate correctness of tool-usage, light-weight tool-using models to test transferability of generated tools), a scaled-up RLVR on tool-design-and-usage will procude capable scientific agents that evolve alongside human knowledge.

REFERENCES

[1] L. Yuan, Y. Chen, X. Wang, Y. R. Fung, H. Peng, and H. Ji, "Craft: Customizing llms by creating and retrieving from specialized toolsets," 2024. [Online]. Available: https://arxiv.org/abs/2309.17428

[2] C. Qian, C. Han, Y. R. Fung, Y. Qin, Z. Liu, and H. Ji, "Creator: Tool creation for disentangling abstract and concrete reasoning of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2305.14318

[3] S. Gao, R. Zhu, P. Sui, Z. Kong, S. Aldogom, Y. Huang, A. Noori, R. Shamji, K. Parvataneni, T. Tsiligkaridis, and M. Zitnik, "Democratizing ai scientists using tooluniverse," 2025. [Online]. Available: https://arxiv.org/abs/2509.23426

[4] T. Cai, X. Wang, T. Ma, X. Chen, and D. Zhou, "Large language models as tool makers," 2024. [Online]. Available: https://arxiv.org/abs/2305.17126

[5] Z. Ma, Z. Huang, J. Liu, M. Wang, H. Zhao, and X. Li, "Automated creation of reusable and diverse toolsets for enhancing llm reasoning," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i23.34664

[6] Z. Wang, D. Fried, and G. Neubig, "Trove: Inducing verifiable and efficient toolboxes for solving programmatic tasks," 2024. [Online]. Available: https://arxiv.org/abs/2401.12869

[7] E. C. Acikgoz, J. Greer, A. Datta, Z. Yang, W. Zeng, O. Elachqar, E. Koukoumidis, D. Hakkani-Tür, and G. Tur, "Can a single model master both multi-turn conversations and tool use? coalm: A unified conversational agentic language model," 2025. [Online]. Available: https://arxiv.org/abs/2502.08820

[8] C. Qian, E. C. Acikgoz, Q. He, H. Wang, X. Chen, D. Hakkani-Tür, G. Tur, and H. Ji, "Toolrl: Reward is all tool learning needs," 2025. [Online]. Available: https://arxiv.org/abs/2504.13958

[9] J. Feng, S. Huang, X. Qu, G. Zhang, Y. Qin, B. Zhong, C. Jiang, J. Chi, and W. Zhong, "Retool: Reinforcement learning for strategic tool use in llms," 2025. [Online]. Available: https://arxiv.org/abs/2504.11536

[10] B. Lyu, Y. Cao, D. Watson-Parris, L. Bergen, T. Berg-Kirkpatrick, and R. Yu, "Adapting while learning: Grounding LLMs for scientific problems with tool usage adaptation," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=owulFly8oQ

[11] A. Miyai, M. Toyooka, T. Otonari, Z. Zhao, and K. Aizawa, "Jr. ai scientist and its risk report: Autonomous scientific exploration from a baseline paper," 2025. [Online]. Available: https://arxiv.org/abs/2511.04583

[12] X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, A. Cohan, D. Greenbaum, Z. Lu, and M. Gerstein, "Risks of AI scientists: prioritizing safeguarding over autonomy," *Nature Communications*, vol. 16, no. 1, p. 8317, Sep. 2025. [Online]. Available: https://doi.org/10.1038/s41467-025-63913-1

[13] J. Wang, Z. Ma, Y. Li, S. Zhang, C. Chen, K. Chen, and X. Le, "Gta: A benchmark for general tool agents," 2024. [Online]. Available: https://arxiv.org/abs/2407.08713

[14] Z. Wang, Q. Chang, H. Patel, S. Biju, C.-E. Wu, Q. Liu, A. Ding, A. Rezazadeh, A. Shah, Y. Bao, and E. Siow, "Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers," 2025. [Online]. Available: https://arxiv.org/abs/2508.20453