Towards Integrated Scientific Discovery Systems

Peter Clark
Allen Institute for AI, Seattle
peterc@allenai.org

Automated/assisted scientific discovery is one of the most exciting emerging areas of AI, fueled by the promise - or at least hope - that neural systems can overcome some of the show-stopping obstacles of the past. At Ai2 we are pursuing this topic, developing an **interactive research agent** in which a user and the system interact through Slack or a Web interface to pursue long-horizon research. In this talk, I'll describe this agent, how it works and is used, and where its strengths and weaknesses are. The agent has access to a number of tools and sub-agents, including for literature search, QA, code execution, and data analysis. First I'll show a (real but cherry-picked) extended example dialog with the agent, in which the user uses these tools to search the literature, identify relevant data, run software experiments, analyze the results, and finally establish whether a hypothesis of interest is true or not, illustrating the vision that we are pursuing.

Following this, I'll describe one of three prototypes we've developed that **performs end-to-end autonomous discovery**, using these and additional tools. The prototype performs a narrow class of computer science research, namely characterizing a LM's ability to perform a task. The input is a LM probing question (e.g., "How well can my LM write stories?"), and the output is a mini technical paper detailing the prototoype's findings. In between, the prototype generates a dataset, runs experiments, analyzes the results, and writes up the findings. This requires not just executing a sequence of tool calls, but also maintaining and updating a representation of the state of research at each step, including the research context, datasets used, hypotheses being considered, and a map of the search space being explored.

I'll then offer some lessons learned from these experiences with end-to-end discovery. First, LMs **do not really know how to do research:** If we simply provide the research tools to the agent, and have it execute a basic "think-then-act" (ReAct) control loop, then the agent does poorly, often going round in circles or getting stuck. Rather, we have had to manually encode a top-level pipeline, which is effective but costly and brittle. In the future, we want to to teach LMs how to do good research, e.g., what are good pipelines and strategies to adopt, and how to manage the "outer loop", namely chaining multiple pipelines together successfully. Second, we find LMs are generally **poor judges of research quality,** making it difficult for LMs to home in on promising research results. While we can use carefully designed quantitative metrics to measure "improved performance", LMs still struggle to assess "novelty" and "impact". Third, discovery involves **systematic search**, a necessary consequence of handling many uncertain decisions during the research process, and something which LMs do not do naturally (being trained for linear action sequences). While we can manually define interesting search spaces, in the future we would like LMs to define search spaces themselves. Fourth, integrated discovery systems are themselves **hard to evaluate**. While new discovery systems appear regularly, it is challenging to know which is "best" and how to learn from these works. As one small contribution to alleviating this problem, we recently released DiscoveryWorld [1], a text-based simulation environment for testing general integrated discovery engines on a variety of discovery quests.

Finally I'll summarize where I see this work going, both within Ai2 and in the wider community, and share my optimism for the Workshop's grand vision of Integrated Approaches to Discovery, and the opportunities ahead

[1] Jansen, P.; Cote, M-A.; Khot, T.; Bransom, E.; Majumder, B.; Dalvi, B.; Tafjord, O.; Clark, P 2024, *DiscoveryWorld: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents*, in NeurIPS'24.
[2] Majumder, B. P.; Surana, H.; Agarwal, D.; Hazra, S.; Sabharwal, A.; and Clark, P. 2024. *Data-driven discovery with large generative models*. In ICML'24.