NovPhy: A Physical Reasoning Benchmark for Open-World AI Systems Author Links Open Overlay Panel (Abstract Reprint)

Vimukthini Pinto[1] , **Chathura Gamage**[1] , **Cheng Xue**[1] , **Peng Zhang**[1] , **Ekaterina Nikonova**[1] , **Matthew Stephenson**[2] and **Jochen Renz**[1]
[1]School of Computing, The Australian National University, Canberra, Australia
[2]College of Science and Engineering, Flinders University, Adelaide, Australia
vimukthini.inguruwattage@anu.edu.au chathura.gamage@flinders.edu.au, cheng.xue@flinders.edu.au, peng.zhang@flinders.edu.au, ekaterina.nikonova@flinders.edu.au, matthew.stephenson@flinders.edu.au, jochen.renz@flinders.edu.au

Abstract Reprint. This is an abstract reprint of a journal article by [Pinto *et al.*, 2024].

Abstract

Due to the emergence of AI systems that interact with the physical environment, there is an increased interest in incorporating physical reasoning capabilities into those AI systems. But is it enough to only have physical reasoning capabilities to operate in a real physical environment? In the real world, we constantly face novel situations we have not encountered before. As humans, we are competent at successfully adapting to those situations. Similarly, an agent needs to have the ability to function under the impact of novelties in order to properly operate in an open-world physical environment. To facilitate the development of such AI systems, we propose a new benchmark, NovPhy, that requires an agent to reason about physical scenarios in the presence of novelties and take actions accordingly. The benchmark consists of tasks that require agents to detect and adapt to novelties in physical scenarios. To create tasks in the benchmark, we develop eight novelties representing a diverse novelty space and apply them to five commonly encountered scenarios in a physical environment, related to applying forces and motions such as rolling, falling, and sliding of objects. According to our benchmark design, we evaluate two capabilities of an agent: the performance on a novelty when it is applied to different physical scenarios and the performance on a physical scenario when different novelties are applied to it. We conduct a thorough evaluation with human players, learning agents, and heuristic agents. Our evaluation shows that humans' performance is far beyond the agents' performance. Some agents, even with good normal task performance, perform significantly worse when there is a novelty, and the agents that can adapt to novelties typically adapt slower than humans. We promote the development of intelligent agents capable of performing at the human level or above when operating in open-world physical environments. Benchmark website: https://github.com/phy-q/novphy

References

[Pinto *et al.*, 2024] Vimukthini Pinto, Chathura Gamage, Cheng Xue, Peng Zhang, Ekaterina Nikonova, Matthew Stephenson, and Jochen Renz. Novphy: A physical reasoning benchmark for open-world ai systems. *Artificial Intelligence*, 336:104198, 2024.