Can we hop in general? A discussion of benchmark selection and design using the Hopper environment

Claas Voelcker
cvoelcker@cs.toronto.edu
University of Toronto
Vector Institute, Toronto

Marcel Hussing
mhussing@seas.upenn.edu
University of Pennsylvania

Eric Eaton
eeaton@seas.upenn.edu
University of Pennsylvania

Abstract

Empirical, benchmark-driven testing is a fundamental paradigm in the current RL community. While using off-the-shelf benchmarks in reinforcement learning (RL) research is a common practice, this choice is rarely discussed. Benchmark choices are often done based on intuitive ideas like “legged robots” or “visual observations”. In this paper, we argue that benchmarking in RL needs to be treated as a scientific discipline itself. To illustrate our point, we present a case study on different variants of the Hopper environment to show that the selection of standard benchmarking suites can drastically change how we judge performance of algorithms. The field does not have a cohesive notion of what the different Hopper environments are representative – they do not even seem to be representative of each other. Our experimental results suggests a larger issue in the deep RL literature: benchmark choices are neither commonly justified, nor does there exist a language that could be used to justify the selection of certain environments. This paper concludes with a discussion of the requirements for proper discussion and evaluations of benchmarks and recommends steps to start a dialogue towards this goal.

1 Introduction

When designing a new algorithm in the wide field of reinforcement learning (RL), a seemingly clear and simple question inevitably arises: *How good is it?*

Theoretical research has made great strides in characterizing the complexity and error bounds of many algorithms under specific assumptions on the MDP structure (Kearns & Singh, 2002; Strehl et al., 2006; Jaksch et al., 2010; Farahmand et al., 2010; Lattimore & Hutter, 2012; Dann & Brunskill, 2015; Osband & Roy, 2016; Azar et al., 2017; Zanette & Brunskill, 2019; Jin et al., 2020b;a; 2021; Domingues et al., 2021). However, once function approximation, large state-action spaces, and exploration are introduced, it is often too difficult to obtain rigorous guarantees without further assumptions (Du et al., 2019; Kane et al., 2022; Golowich et al., 2024). In lieu of mathematical yardsticks, the empirical RL community has used benchmarks and competitive testing to obtain performance estimates of proposed algorithms (Bellemare et al., 2013; Brockman et al., 2016; Tunyasuvunakool et al., 2020). While the statistical validity of empirical comparisons have received attention (Henderson et al., 2018; Agarwal et al., 2021; Patterson et al., 2023), another important problem is less discussed: *Are our benchmarks representative of a wider set of problems of interest?*

This paper is a play in two parts. In act one, we provide a technical evaluation. We showcase that picking among two different variants of similar benchmarks, we are unable to replicate algorithm evaluation. For this, we investigate four algorithms: Soft-actor critic (SAC) (Haarnoja et al., 2018), Model-based Policy Optimization (MBPO) (Janner et al., 2019), Aligned Latent Models (ALM) (Ghugare et al., 2023), and Diversity is All You Need (DIAYN) (Eysenbach et al., 2019), chosen to account for the variety of RL paradigms (model-free, model-based, and reward-free). We highlight that derived insights do not generalize across the different implementations for all these algorithms. This shows the field does not have a cohesive notion *of what* the different Hopper environments are representative—they do not even seem to be representative of each other.

In the second act, we form a position statement by reflecting on the outcome of the technical section. We argue that our experiments necessitate a reorientation of the community on the role and selection of benchmarks. In this, we aim to widen the frame of reinforcement learning research to include research on the benchmarks themselves. The intuitive idea of a "Hopper" does evidently not capture the relevant aspects and difficulties of the benchmark. Without evaluating its benchmarks, the RL community is unable to fully quantify whether there is genuine progress towards the goal of general learning agents. We summarize the overarching claim of these two acts in a single statement:

Benchmark selection impacts the evaluation of algorithms, but in the past it has largely been neglected as a first-class problem. This necessitates a re-evaluation of how we describe, compare, and design our RL evaluation platforms.

2 Background

Before we discuss whether commonly used benchmarks are adequate, we require at least a rough definition of the term "benchmark". Although benchmarks play a vital role in the development of machine learning (Deng et al., 2009; Raji et al., 2021), there is little formal study of their design and role, especially in the context of RL. Thus, we synthesize an informal definition from a dictionary (dic) and prior work asking a similar question to ours in image classification (Raji et al., 2021).

Definition 2.1 (Benchmark) *A benchmark is a software library or dataset together with community resources and guidelines specifying its standardized usage. Any benchmark comes with predefined, rankable performance metrics meant to enable algorithm comparison.*

To be a little more meticulous about this definition, we can try to draw from other fields that have treated benchmarking as an area of study. In performance engineering, von Kistowski et al. (2015) define various lower-level characteristics that are helpful for us to understand what a benchmark is. It should be relevant to the behavior of interest, reproducible, fair, verifiable, and easily usable. While all of these play a crucial role also for the RL community, our focus will be on the first point.

To understand relevance, we differentiate benchmarks from domain-specific test environments. An example of the latter would be a specific robot simulator for platform for which an engineer requires a control algorithm. The engineer might use this simulator of a physical platform, carefully test the difference between the simulation and the real environment, and then design and tune an RL algorithm to produce a controller which is executed in the real environment. A benchmark however is used not to obtain an instantiation of a policy, but rather to test whether a method to obtain policies performs well on the specified metric. The important differentiating factor is that an RL researcher does not necessarily care about an agent that, for example, plays Atari games well. Atari games serve as a proxy for an actual problem of interest interest which may be hard to specify.

Given that the performance on the specific metric is not the main goal, we instead hope that it is *correlated* to a problem *representative* of a wider class of interesting scenarios within which our finding generalize. This is is rarely explicitly assessed and there are philosophical difficulties in establishing what qualifies as *representative* (Chollet, 2019). Notably, in fields like image classification and natural language processing, discussing if and how benchmarks actually represent problems of interest is an active subfield (Scheuerman et al., 2021), yet this discussion is mostly absent in the RL literature. We defer a further discussion of the nuances of this question to the second part of the paper, and simply posit the following, somewhat testable core requirement for now:

If an algorithm performs well on a benchmark environment,
it should also do well on other environments representing similar objectives.

3 Empirical case study: The Hopper environment

For continuous control, a standard testing scenario includes simplified locomotion benchmarks. In these, the goal is to make various instantiations of legged robots move forward. There are two commonly used variants of this benchmark, the OpenAI Gym (Brockman et al., 2016) and the DeepMind Control (Tunyasuvunakool et al., 2020) suites. Both are built on top of the MuJoCo physics simulator (Todorov et al., 2012) and contain similar robots and tasks, making them well-suited for testing our previous requirement. Our focus will specifically be the Hopper robot, whose goal it is to hop forward without falling over. Paraphrasing our requirement, one might be tempted to say: If it looks like a Hopper and hops like a Hopper, then it better be able to hop everywhere.

To verify whether algorithms are able to make the Hopper hop, we selected four commonly used algorithms. We chose SAC (Haarnoja et al., 2018) as one of the most popular model-free RL algorithms, MBPO (Janner et al., 2019) and ALM (Ghugare et al., 2023) as representatives of the model-based paradigm, and DIAYN (Eysenbach et al., 2019), a reward-free RL approach. For details on our implementation choices we refer to Appendix A. Each algorithm was originally published using the OpenAI gym variant of the benchmark. We first discuss the major differences between the two implementations and then present a detailed comparison of algorithm performance.

3.1 Comparing the benchmark specifications

The Hopper environment is, to the best of our knowledge, first mentioned in Erez et al. (2011) and it reappears in Schulman et al. (2015) and Lillicrap et al. (2016). In the current Gym version (Towers et al., 2023), the task description in the documentation (May 2024) is *"The Hopper is a two-dimensional one-legged figure that consist of four main body parts - the torso at the top, the thigh in the middle, the leg in the bottom, and a single foot on which the entire body rests. The goal is to make hops that move in the forward (right) direction by applying torques on the three hinges connecting the four body parts."* (hop) The reward model consists of three terms: a bonus for remaining *healthy*, which is defined with a set of constraints to ensure that the Hopper is roughly upright, a bonus for achieving a forward velocity, and a small regularization cost for large torques. The episode is also terminated if the Hopper becomes "unhealthy", meaning it has toppled over.

The other variant of the Hopper benchmark is a reimplementation of the environment as part of the DeepMind Control Suite of environments (Tunyasuvunakool et al., 2020) based on (Lillicrap et al., 2016). In this variant, there are three important changes. The Hopper's torso has one additional link, which also adds another controllable joint. The reward is designed to be strictly contained within the interval $[0, 1]$, making comparisons of algorithms simpler as the maximum cumulative return can be computed from the episode length. Finally, the DMC suite does not use early termination of unhealthy environments, trajectories are truncated after 1000 steps by default. These design choices lead to an effectively sparser reward in the DMC Hopper environment, as the Hopper obtains 0 reward when toppling over, but the episode is not reset.

Both variants of the Hopper environment ostensibly represent a very similar robot and task but we will demonstrate that it does not fulfill our core requirement: neither algorithmic performance evaluation nor qualitative claims about the usefulness of the chosen algorithms generalize between these two environments. Note that we do not test whether a policy is transferable between simulators, but whether the same algorithm is able to produce a policy on both environments.

3.2 Reward-based reinforcement learning

We compare the four previously mentioned algorithms in the Hopper environment using three different instantiations: the standard Gym version of the benchmark, a modified Gym version without forced termination (to test whether the gap mostly stems from early termination), and the DM Control version. We report our results in Figure 1.

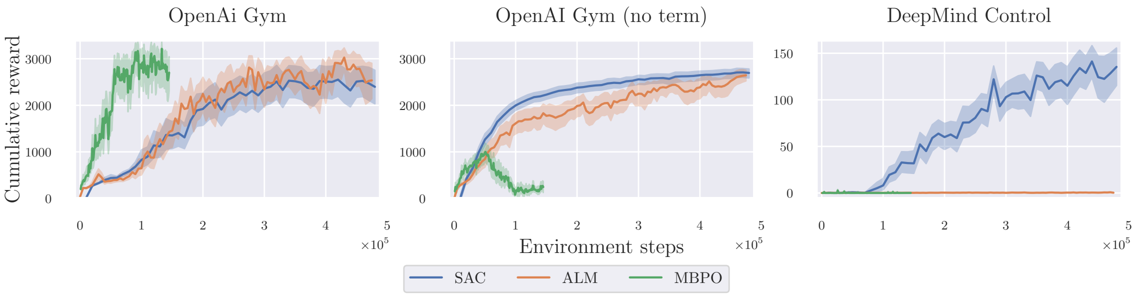

Figure 1: Performance evaluation on the Hopper environment variations. Shaded area denotes a bootstrapped 95% confidence interval of the mean at 95 across 30 seeds with 5000 resamples.

SAC is able to solve OpenAI Gym variant of the benchmark, but struggles with the DMC version. To the best of our knowledge, some recent algorithms achieve at most roughly 500 reward points on this benchmark (D'Oro et al., 2023; Hussing et al., 2024; Hansen et al., 2024). However, both model-based approaches do not achieve any reward on the DMC variant, even though they achieve on par or better results on the OpenAI Gym version. This means that depending on which variant of the benchmark is used, neither the absolute performance nor the relative ranking of the algorithms stays intact, a feature that other benchmarks such as ImageNet do possess (Recht et al., 2019).

The hypothesis that this effect can be attributed mostly to early termination cannot be verified. While MBPO struggles with the absence of termination, both SAC and ALM obtain similar returns to the early terminated variant. Further investigation on the performance of model-based algorithms in environments without early termination seems pertinent in this light. For completeness sake, it is important to note that not all model-based approaches fail on the DMC Hopper task. Algorithms such as TD-MPC2 achieve strong performance even without termination (Hansen et al., 2024).

3.3 Reward-free reinforcement learning

For reward-free RL, we present a qualitative evaluation. First, we note that the original work presenting the DIAYN algorithm (Eysenbach et al., 2019) tests their approach on the Hopper variant with early termination. Their results are presented at this link: sites.google.com/view/diayn/hopper. In their evaluation, the Hopper performs a variety of different dynamic skills, which can be used to obtain policies for downstream tasks.

We repeated the same experiment now on the DMC variant, without early termination. We visualize the end states of the trajectories in Figure 2, with the full trajectories visualized in the appendix. Without termination, the optimization target of DIAYN does not incentivize learning dynamic skills, but rather distinct static poses, as these are simple to distinguish for the discriminator. As most common definitions of "skill" imply a dynamic motion, not a static contortion, we highlight that the emergence of skills with DIAYN seems to depend on properties of the *testing environment* as much as on the proposed *algorithm*. Thus, if another variant of the benchmark, had been chosen, the analysis of the algorithm would likely have been different. This again highlights our core claim.

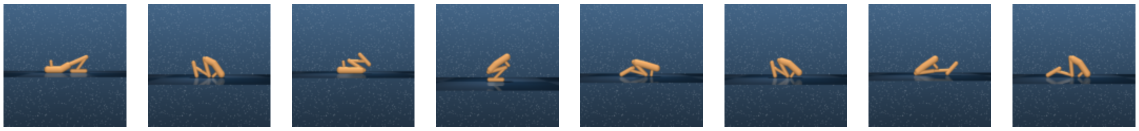

Figure 2: Final states for 8 out of 20 "skills" learnt by the DIAYN algorithm on DMC Hopper. In all cases the Hopper immediately moves towards the final configuration without displaying the dynamic skills reported in the original paper.

4 How do we describe and evaluate benchmarks?

In some sense, the previous section can be narrowly read as a simple comparison of a single environment out of set of two common test benchmarks. Finding that these two benchmarks behave differently, especially when realizing that not even the observation or action spaces are identical, might not be too surprising. However, dismissing the experiments ignores the role and use of benchmarks in the wider context of the RL literature. The fact that the observation and action spaces are different *should not matter to this degree.* If you agree with our initial requirement, namely that algorithms should work across benchmarks that represent the same agent and have the same objective, we may have convinced you at this point that something is wrong.

If a benchmark is supposed to be representative of a wider set of algorithmic problems, we must begin with investigating the goal of building and using the benchmark. The question that then demands an answer is: What is the ultimate purpose of reinforcement learning? The broad purpose of AI is often stated as the development and understanding of general intelligence, as vague as that might be. If this is indeed the intended purpose, the fact that our algorithms performs so differently on benchmarks with arguably similar objectives, is —at least to us— somewhat unsettling.

4.1 Evaluating the purpose of reinforcement learning

Reinforcement learning as a field studies methods to compute optimal policies (or related quantities e.g. value functions) in Markov Decision Processes. The fundamental object here is not derived by a real-world application such as with image classification, but is instead a formal, mathematical object that can be arbitrarily complex. The generality of the MDP is often seen as a strength of the RL paradigm, as many problems which would intuitively require intelligence can be phrased as a specific variant of the general framework. However, this generality also causes problems.

If the purpose of our field was simply the mathematical study of MDPs, then empirical research is not needed at all. However, finding efficient algorithms to solve *any* MDP is likely an intractable proposal. Instead, we might want efficient solutions to many real world problems by understanding their properties (Farahmand, 2011). Approaching the overarching purpose from the empirical side as well to establish relevant (sub-)problems on the path towards general intelligence is then indeed important. We invite the community to approach the resulting question, even though it is difficult: What specific (sub-)problems are important to solve with our algorithms and how do we measure whether our benchmarks properly capture these problems?

We acknowledge the main challenge with our ultimate purpose: Verifying that a benchmark meaningfully represents general intelligence is incredibly difficult. Thus, we argue that RL benchmarking needs to be treated as a scientific discipline itself. This is not a criticism of the hard work that has already been put into developing novel RL benchmark environments. We value the contributions of such work as establishing new test beds is one way to address this difficult problem. To acknowledge a sampling of such efforts, we briefly survey related work on benchmarks in Appendix C. Rather, similar to recent efforts in evaluating Large Language Models, we argue that the RL community needs to start thinking about how to describe its benchmarks and motivate their selection. In addition to more diverse and interesting benchmarks, we call for expanding the frame of RL research to include assessing the benchmarks themselves.

4.2 Towards effective benchmarks: Developing a common language

One of the first steps in this endeavor we propose is the development of a common language for benchmarks. We need to establish an understanding of important concepts, such as goals, properties, and measurable quantities that are often used without proper specification. In this work, we have already introduced some nomenclature but we would like to emphasize that it might by no means be perfect or the standard from here on out. It is an attempt to start a conversation around a topic we believe to be quite difficult to grasp. We have established the term *purpose* for the overarching objective of the community and *problems* as concrete challenges on the path towards this goal.

We use the term *property* to describe the intuitive categories that researchers have used to describe and group benchmarks, such as “legged robot” or “visual observation”. However, our work hopefully convinces you that these are not necessarily useful! Purely intuitive notions of “properties” such as “Hopper” cannot be used without establishing and testing whether they represent meaningful groupings with regard to problems of interest. We do not know how to do this in all generality but at least we have established that such tests can be attempted. While some attempts have been made to define properties of MDPs more rigorously (Osband et al., 2020), to the best of our knowledge no work has attempted to formalize and study these properties in most established benchmarks.

To develop such properties, we require measurable quantities to group and compare environments, and to test whether they are meaningful for the problem they were defined for. One positive example that studies properties of benchmarks is Laidlaw et al. (2023). It introduces the “effective horizon” as a testable quantity of exploration difficulty in the Atari games. Another example from the empirical literature is Machado et al. (2018), who investigate common usage patterns of the ALE suite. Importantly, these quantities need to be computable across arbitrary benchmarks to allow us to draw conclusion across test-beds. Moreover, they might be highly entangled with each other and getting to a state of disentanglement will require nuanced treatment and a continually improving research process. For example, when measuring the exploration difficulty in a pixel-based environment, the visual representation problem could be a confounding factor. Extending and verifying these quantities should be a celebrated contribution to the literature at relevant venues.

Finally, it is necessary for future RL papers to discuss how the chosen benchmarks impact algorithmic design choices. For example, our experiments show that for DIAYN, early termination is not merely an incidental property of the benchmark that is external to and separate from the algorithm. Instead, it is important so that the algorithm behaves as intended. By not discussing such properties directly, we structurally blind ourselves to understanding their impact. An outcome of the study of benchmarks should be the ability to inform a conscious choice of the appropriate environments.

4.3 A peek at other fields of science

Purpose, goals, properties and measurable quantities have a somewhat circular relationship as one needs to define a purpose to define a goal and corresponding properties and quantities but we require disentangled quantities to reason which goals are achievable. At this point we would like to draw a parallel to physical sciences. In physics, science often starts out with a first stab to characterize a system. Isaac Newton defined a “Force” as a measurable quantity consisting of mass and acceleration. For a long time, this was sufficient to explain various physical behaviors. With the study of more complex systems such as atomic system, these definitions needed to be revisited. We believe that a similar, open-ended research process will be necessary to characterize benchmarks.

Thus, we should allow researchers that attempt to make progress towards this end to publish their work even if it may not capture every single facet of the problem. Similar to how we often take small steps to develop solutions in the algorithm realm, we should allow ourselves to make incremental but consistent progress on the relevance of benchmarks. Many other fields have worked on the question of how to measure important quantities of interest (take for example the enormous difficulty of establishing a meaningful measure of human intelligence), and how to test whether novel approaches make meaningful progress towards them. These can serve as inspiration for us, for example reviewing the thorough and rigorous testing standards of engineering fields such as material science.

5 Conclusion

We have questioned the motivation for benchmark usage in RL and ask the community to consider a meta-level question that will help inform more conscious choices around which benchmarks are relevant for what. More work is needed to properly characterize what our benchmarks *represent*.

What is the ultimate purpose and what are the underlying goals for the RL community?
And how do we capture its various aspects in our benchmark environments in a testable manner?

Acknowledgments

We thanks the anonymous reviewer for their help, especially with framing the abstract to clarify the positioning of the paper.

This research was partially supported by the Army Research Office under MURI award W911NF20-1-0080, the DARPA Triage Challenge under award HR001123S0011, and by the University of Pennsylvania ASSET center. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA, the Army, or the US government.

References

"benchmark." merriam-webster.com. https://www.merriam-webster.com. Accessed: 2024-05-24.

"hopper." gymnasium documentation. https://gymnasium.farama.org/environments/mujoco/hopper/. Accessed: 2024-05-24.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *9th International Conference on Learning Representations (ICLR-21)*, 2021.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 06–11 Aug 2017.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research (JAIR)*, 47:253–279, 2013.

Clément Bonnet, Daniel Luo, Donal Byrne, Shikha Surana, Sasha Abramowitz, Paul Duckworth, Vincent Coyette, Laurence I. Midgley, Elshadai Tegegn, Tristan Kalloniatis, Omayma Mahjoub, Matthew Macfarlane, Andries P. Smit, Nathan Grinsztajn, Raphael Boige, Cemlyn N. Waters, Mohamed A. Mimouni, Ulrich A. Mbou Sob, Ruan de Kock, Siddarth Singh, Daniel Furelos-Blanco, Victor Le, Arnu Pretorius, and Alexandre Laterre. Jumanji: a diverse suite of scalable reinforcement learning environments in jax, 2024. URL https://arxiv.org/abs/2306.09884.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *CoRR*, abs/1606.01540, 2016.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First steps towards grounded language learning with a human in the loop. In *7th International Conference on Learning Representations (ICLR-19)*, 2019.

François Chollet. On the measure of intelligence, 2019.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532, 2018.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2818–2826, Cambridge, MA, USA, 2015. MIT Press.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OpC-9aBBVJe.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

T. Eimer, A. Biedenkapp, M. Reimer, S. Adriaensen, F. Hutter, and M. Lindauer. Dacbench: A benchmark library for dynamic algorithm configuration. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21)*. ijcai.org, August 2021.

Tom Erez, Yuval Tassa, and Emanuel Todorov. Infinite-horizon model predictive control for periodic tasks with contacts. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2011. doi: 10.15607/RSS.2011.VII.010.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/65cc2c8205a05d7379fa3a6386f710e1-Paper.pdf.

Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Russ Salakhutdinov. Simplifying model-based RL: Learning representations, latent-space models, and policies with one objective. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MQcmfgRxf7a.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploration is harder than prediction: Cryptographically separating reinforcement learning from supervised learning. *arXiv preprint arXiv:2404.03774*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Peter Henderson, Wei-Di Chang, Florian Shkurti, Johanna Hansen, David Meger, and Gregory Dudek. Benchmark environments for multitask learning in continuous domains. *ICML Lifelong Learning: A Reinforcement Learning Approach Workshop*, 2017.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Marcel Hussing, Claas Voelcker, Igor Gilitschenski, Amir-massoud Farahmand, and Eric Eaton. Dissecting deep rl with high update ratios: Combatting value overestimation and divergence. *arXiv preprint arXiv:2403.05996*, 2024.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 13–18 Jul 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13406–13418. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6f5e4e86a87220e5d361ad82f1ebc335-Paper.pdf.

Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1282–1302. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/kane22a.html.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49:209–232, 2002.

Michal Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaskowski. Vizdoom: A doom-based AI research platform for visual reinforcement learning. *CoRR*, abs/1605.02097, 2016.

Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4501–4510, Apr. 2020.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack Learning Environment. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Cassidy Laidlaw, Stuart J Russell, and Anca Dragan. Bridging rl theory and practice with the effective horizon. *Advances in Neural Information Processing Systems*, 36:58953–59007, 2023.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, abs/1908.09453, 2019.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *Algorithmic Learning Theory*, pp. 320–334, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34106-9.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8c3c666820ea055a77726d66fc7d447f-Paper-Datasets_and_Benchmarks.pdf.

Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. *ACM International Conference on AI in Finance (ICAIF)*, 2021.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Jorge A. Mendez, Marcel Hussing, Meghna Gummadi, and Eric Eaton. CompoSuite: A compositional reinforcement learning benchmark. In *1st Conference on Lifelong Learning Agents (CoLLAs-22)*, 2022.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/eda80a3d5b344bc40f3bc04f65b7a357-Paper-round2.pdf.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning, 2016.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, Benjamin Van Roy, Richard Sutton, David Silver, and Hado Van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygf-kSYwH.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*, 2023.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile

Gervet, Vincent-Pierre Berges, John M Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*, 2024.

Deborah Raji, Remi Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 881–888, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-12)*, pp. 5026–5033, 2012.

Tristan Tomilin, Meng Fang, Yudi Zhang, and Mykola Pechenizkiy. Coom: A game benchmark for continual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023. URL https://zenodo.org/record/8127025.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John P. Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy P. Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. StarCraft II: A new challenge for reinforcement learning. 2017.

Jóakim von Kistowski, Jeremy A. Arnold, Karl Huppler, Klaus-Dieter Lange, John L. Henning, and Paul Cao. How to build a benchmark. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, ICPE '15, pp. 333–336, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450332484.

Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual World: A robotic benchmark for continual reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS-21)*, 2021.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the 3rd Conference on Robot Learning (CoRL-19)*, 2019.

Zhaocong Yuan, Adam W. Hall, Siqi Zhou, Lukas Brunke, Melissa Greeff, Jacopo Panerati, and Angela P. Schoellig. Safe-control-gym: A unified benchmark suite for safe learning-based control and reinforcement learning in robotics. *IEEE Robotics and Automation Letters*, 7(4):11142–11149, 2022. doi: 10.1109/LRA.2022.3196132.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7304–7312. PMLR, 09–15 Jun 2019.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

A Implementation details of algorithms

Algorithm	Code base
SAC	https://github.com/proceduralia/high_replay_ratio_continuous_control
MBPO	https://github.com/facebookresearch/mbrl-lib
ALM	https://github.com/RajGhugare19/alm
DIAYN	Personal reimplementation based on https://github.com/proceduralia/high_replay_ratio_continuous_control

Table 1: Code sources for tested algorithms

The source code repositories for all algorithms are presented in Table 1. We aimed to make minimal changes to the algorithms and hyperparameters, to stay close to the way that they have been evaluated in previous work.

We used all algorithms with the same hyperparameters across all environments, taken from the original papers proposing each algorithm or the off-the-shelf reference implementations. We note that all hyperparameters are similar to each other, as all algorithms use SAC as the actor-critic component, and even with some moderate hyperparameter tuning, we were unable to elicit good performance from MBPO and ALM on DMC. We also note that while MBPO is seemingly trained for significantly fewer steps, this is in line with the original paper, as MBPO uses a significantly higher update ratio.

All of our chosen algorithms were originally implmented on the OpenAI Gym benchmark variant. This was a conscious choice, for two reasons. First, the OpenAI Gym version is the more popular benchmark alternative, which means more algorithm in general are tested on it. Second, the existence of early termination requires explicit handling in the code. To minimize the amount of changes to the code we were required to make, it was useful to chose algorithm implementations that already handle this correctly. An algorithm like T-MPC2 (Hansen et al., 2024) would have required potentially substantial changes to be useable on OpenAI Gym.

B DIAYN sequences

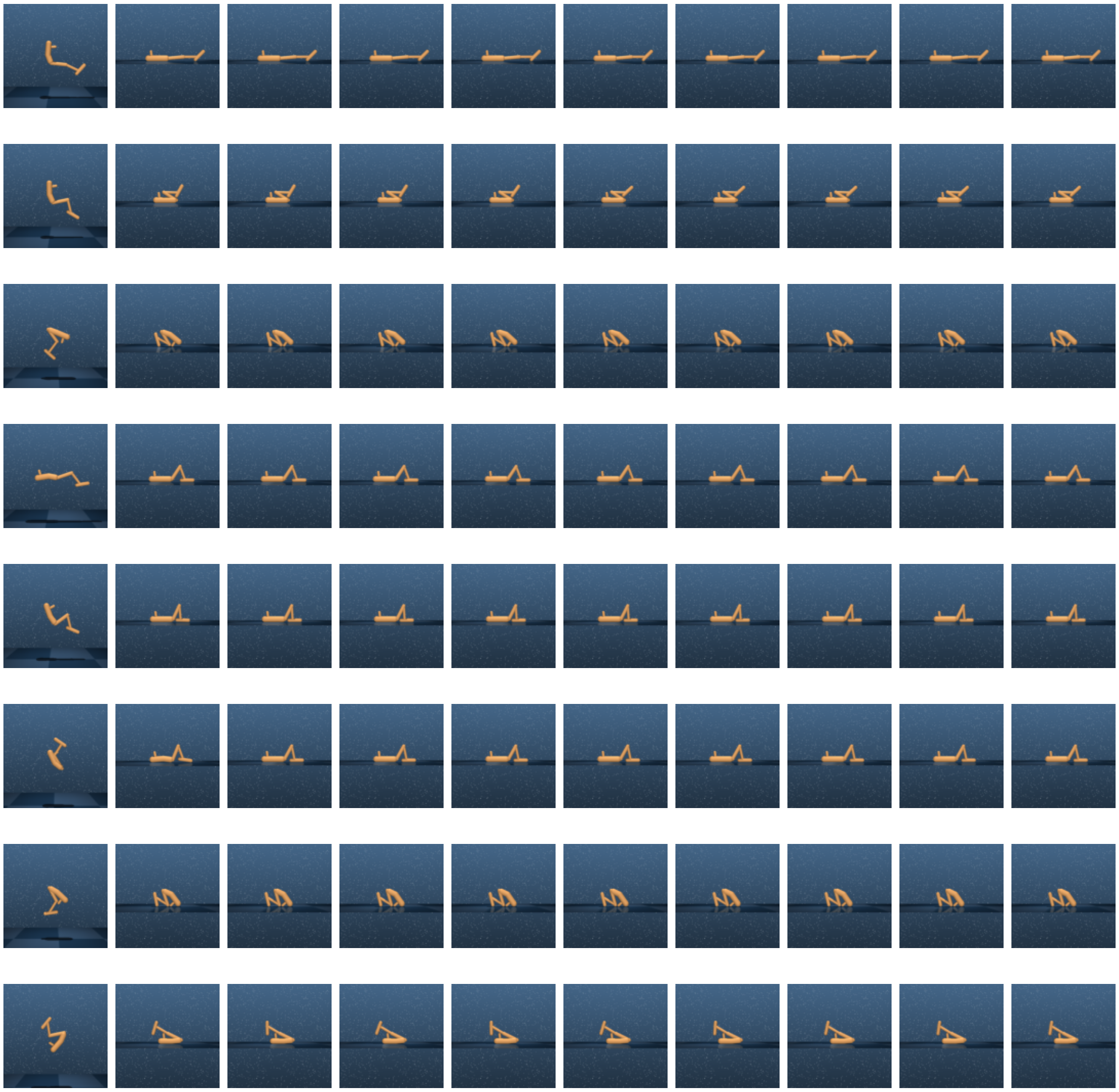

Figure 3: Full visualization of 8 out of 20 skills from one run of DIAYN. Each row is one sequence from the execution of a skill, with each column one frame every 100 environment time steps. For almost all skills we observe that the Hopper maintains a static pose after the first 50-100 time-steps, which is a very different behavior from the one reported in the original paper.
The poses are mostly distinct, which does optimize the discriminative objective of the algorithm.

C Additional related work on benchmarks

The development of new testbeds has played a significant role in the area of reinforcement learning. The following is a non-exhaustive list of highly valuable benchmark contributions to the field of RL.

A classical application for benchmarking in RL has always been video games due to the planning complexity and our ability to quickly simulate them (Bellemare et al., 2013; Brockman et al., 2016; Kempka et al., 2016; Vinyals et al., 2017; Lanctot et al., 2019; Kurach et al., 2020). When moving to continuous control, benchmarks are often inspired by robotics application. This includes for instance the set of DM control tasks used in this work (Tunyasuvunakool et al., 2020). Yet, there are many other attempts to capture the difficulties of continuous control in single task learning (Zhu et al., 2020; James et al., 2020). To build challenges beyond the single-task realm, others have tried to capture modes beyong single task learning. There exist benchmarks for multi-task (Henderson et al., 2017), meta (Yu et al., 2019), continual (Wołczyk et al., 2021; Tomilin et al., 2023; Liu et al., 2023), compositional (Mendez et al., 2022), skill (Mu et al., 2021) and causal (Ahmed et al., 2021) learning. Further, researchers have integrated other challenges posed by different modalities such as text (Côté et al., 2018; Chevalier-Boisvert et al., 2019; Küttler et al., 2020) or visual generalization (Cobbe et al., 2019) as well as challenges posed by human-robot collaboration (Puig et al., 2024) into the their benchmarks. There are also benchmarks that are more application-based and try to solve automatic algorithm configuration (Eimer et al., 2021), combinatorial problems (Bonnet et al., 2024) or financial trading (Liu et al., 2021). Finally, there are benchmarks which focus on tasks other than reward maximization, such as safe control (Ray et al., 2019; Yuan et al., 2022).