A2Perf: Benchmarking Autonomous Agents End-to-End in Realistic Domains

Ikechukwu Uchendu [1 *] **Jason Jabbour** [1] **Korneel Van den Berghe** [2] **Joel Runevic** [1] **Matthew Stewart** [1] **Jeffrey Ma** [1] **Srivatsan Krishnan** [1] **Izzeddin Gur** [3] **Austin Huang** [3] **Colton Bishop** [3] **Paige Bailey** [3] **Joe Wenjie Jiang** [3] **Ebrahim M. Songhori** [3] **Sergio Guadarrama** [3] **Jie Tan** [3] **Jordan K. Terry** [4] **Aleksandra Faust** [3 ‡] **Vijay Janapa Reddi** [1 † ‡]

Abstract

Autonomous agents and systems cover a number of application areas, from robotics and digital assistants to combinatorial optimization, all sharing common, unresolved research challenges. It is not sufficient for agents to merely solve a given task; they must generalize to out-of-distribution tasks, perform reliably, and use hardware resources efficiently during training and on-device deployment, among other requirements. Several classes of methods, such as reinforcement learning and imitation learning, are commonly used to tackle these problems, each with different trade-offs. However, there is a lack of benchmarking suites that define the environments, datasets, and metrics which can be used to provide a meaningful way for the community to compare progress on applying these methods to real-world problems. We introduce A2Perf —a benchmarking suite including three environments that closely resemble real-world domains: computer chip floorplanning, web navigation, and quadruped locomotion. A2Perf provides metrics that track task performance, generalization, system resource efficiency, and reliability, which are all critical to real-world applications. In addition, we propose a data cost metric to account for the cost incurred acquiring offline data for imitation learning, reinforcement learning, and hybrid algorithms, which allows us to better compare these approaches. As an open-source and extendable benchmark, A2Perf is designed to remain accessible, documented, up-to-date, and useful to the research community over the long term.

[*]Work done as a student researcher at Google DeepMind. [†]Work done as a visiting researcher at Google DeepMind. [‡]Equal advising. [1]Harvard University [2]Delft University of Technology [3]Google DeepMind [4]Farama Foundation. Correspondence to: Ikechukwu Uchendu <iuchendu1@gmail.com>.

Proceedings of the ICML 2025 Workshop on Championing Open-source Development in Machine Learning (CODEML '25). Copyright 2025 by the author(s).

1. Introduction

Autonomous agents observe their environment, make decisions, and perform tasks with minimal human interference (Sutton & Barto, 2018). In this work, we generally focus on autonomous agents trained with reinforcement learning (RL) and imitation learning (IL) techniques, as opposed to agents based on foundation models (Bommasani et al., 2021). While these RL and IL agents have been successfully evaluated across a wide range of application domains, developing algorithms for their deployment in real-world scenarios presents significant challenges (Dulac-Arnold et al., 2021). These challenges include dealing with high-dimensional state and action spaces, partial observability, non-stationarity, sparse rewards, and the need for safety constraints. Furthermore, real-world environments often have multiple objectives, require sample efficiency, and necessitate robust and explainable decision-making. Addressing these challenges is crucial for productionizing reinforcement learning algorithms to real-world problems.

To enable researchers to develop algorithms with real-world deployment considerations in mind, there is a need for benchmarks that incorporate practical metrics. These include metrics such as the compute required for training and inference, wall-clock time, and effort expended on data collection. While there are existing benchmarks for autonomous agents (Guss et al., 2019; Yu et al., 2020; Kempka et al., 2016; Bellemare et al., 2013; Chevalier-Boisvert et al., 2023; Tassa et al., 2018), most only evaluate an agent's raw performance on the same task on which it was trained, without considering numerous other metrics that matter in real-world production training and deployment scenarios.

In this paper, we introduce A2Perf[1], a benchmarking framework that aims to bridge the gap between algorithms research and real-world applications by providing a comprehensive evaluation platform for autonomous agents, thereby expanding the applicability of reinforcement learning to a wide range of practical domains. In addition, it comes equipped with a critical set of metrics for fair assessment.

[1]The project website and documentation can be found at `https://a2perf.farama.org/`

Benchmark	Metrics				Real-World Tasks	Offline Datasets
	Generalization	System	Data Cost	Reliability		
A2Perf	✓	✓	✓	✓	✓	✓
D5RL (Rafailov et al., 2024)	✓	✗	✗	✗	✓	✓
NeoRL (Qin et al., 2022)	✗	✗	✗	✗	✓	✓
OGBench (Park et al., 2024)	✓	✗	✗	✗	✓	✓
Meta-World (Yu et al., 2020)	✓	✗	✗	✗	✓	✗
DM Control (Tassa et al., 2018)	✗	✗	✗	✗	✗	✗
Jumanji (Bonnet et al., 2023)	✓	✗	✗	✗	✓	✗
DSRL (Liu et al., 2023)	✗	✗	✗	✗	✓	✓
Safety Gym (Ji et al., 2023)	✗	✗	✗	✓	✓	✗
ALE (Bellemare et al., 2013)	✗	✗	✗	✗	✗	✗
MineRL (Guss et al., 2019)	✓	✗	✗	✗	✗	✓
Loon Benchmark (Greaves et al., 2021)	✓	✗	✗	✗	✓	✓

Table 1: A2Perf compared to existing benchmarks that evaluate autonomous agents. Checkmarks (✓) indicate the presence of a feature or metric, while crosses (✗) denote its absence. A2Perf distinguishes itself by including metrics for generalization, system resource efficiency, data cost, and reliability, in addition to providing real-world tasks and offline datasets. Here, real-world tasks refer to those that are often performed in industrial or consumer contexts. The selected domains in A2Perf are designed to closely mirror real-world challenges, ensuring the relevance and transferability of the benchmark results to practical applications.

A2Perf incorporates three challenging domains based on prior work (Coumans, 2023; Mirhoseini et al., 2021; Gur et al., 2021) that closely mirror scenarios that have been demonstrated in the real world: computer chip-floorplanning, website form-filling and navigation, and quadruped locomotion. In addition, these domains were chosen because they inherently exhibit a small Sim2Real gap. The computer chip floorplanning domain (Mirhoseini et al., 2020; 2021) was used to help create an iteration of Google's tensor processing unit[2], where the agent optimizes the layout of chip components. In the website form-filling and navigation domain (Gur et al., 2018; 2021), agents autonomously navigate and interact with websites in a Google Chrome[3] browser, making it identical to real-world web navigation. The quadruped locomotion domain (Peng et al., 2020) has demonstrated successful transfer of learned walking gaits to the Unitree Laikago[4] robot.

Furthermore, to address the metrics gap, A2Perf provides an open-source benchmarking suite that evaluates agents across four key metric categories: (1) data cost, which quantifies the effort required to gather training data for imitation learning, (2) application performance, relating to the quality of the agent's task-specific execution, and its ability to generalize to tasks that it was not explicitly trained to perform; (3) system resource efficiency, focusing on the hardware resources used during training and inference; and (4) reliability, denoting the consistency of an agent's performance over training and inference. While three domains and four classes of metrics are currently available, A2Perf allows for straightforward expansion to benchmark on custom domains and for custom metrics.

2. Metrics for Real-World Evaluation

2.1. Data Cost

Comparing agent performance trained using different approaches is challenging but important to gain a holistic picture of the costs and trade-offs involved. IL methods may be more sample efficient than RL methods, as they do not need to interact with the environment online. However, this perspective overlooks the *effort* required to collect demonstration data used for IL.

To facilitate fair comparisons between these approaches, we propose the **training sample cost** metric, which quantifies the effort required to obtain offline datasets used by the agent. In this context, we denote the training sample cost of an offline dataset D as C_D. An agent that uses samples from datasets $D_1, D_2, \ldots, D_K$ will incur a total training sample cost of $\texttt{Training Sample Cost} = \sum_{i=1}^{K} C_{D_i}$. The datasets D_i could be of different *expertise* levels, meaning they contain demonstrations from agents or humans with varying levels of task proficiency.

The training sample cost can be measured with any metric that meaningfully represents the effort required to generate

[2]History of the Tensor Processing Unit: https://shorturl.at/Bo71S

[3]Google Chrome Browser: https://www.google.com/chrome/

[4]Unitree Laikago: https://shorturl.at/FD6uP

samples for imitation learning. For example, the cost could be expressed in terms of money spent on human labor or computational resources, hours invested in collecting the data, or any other relevant metric. The choice of metric may depend on the specific application and the type of data being collected since training samples can originate from a variety of sources, such as human operators (Mandlekar et al., 2020), pre-existing policies (Hester et al., 2018), or logged experiences from different agents (Fujimoto et al., 2019; Kostrikov et al., 2021).

In A2Perf, we adopt a simplified approach by focusing on datasets generated solely from RL policies, using energy consumption as our training sample cost metric. This design choice enables systematic evaluation while avoiding the complexities of collecting and pricing human demonstrations. Specifically, we define the training sample cost, C_D, of a dataset D as the average energy consumed to train the policies that are used to generate the dataset D. This can be expressed as:

$$C_D = \frac{1}{|\Pi_D|} \sum_{\pi \in \Pi_D} E_{\text{train}}(\pi) \tag{1}$$

where Π_D is the set of policies used to generate the dataset D, $|\Pi_D|$ denotes the number of policies in this set, and $E_{\text{train}}(\pi)$ represents the energy consumed to train the policy π. As we strive for more equitable comparisons between approaches to training autonomous agents, we urge the research community to consider the cost of acquiring training data. To this end, we release datasets for each domain and task in A2Perf, along with their associated training sample costs. While the specific expertise levels may vary across domains and tasks, we generally consider three categories: `novice`, `intermediate`, and `expert`. See Appendix H for the dataset collection procedure and Appendix I for details on the dataset format.

2.2. System Performance

System metrics provide insight into the feasibility of deploying autonomous agents, particularly considering the scaling demands on energy and data efficiency (Frey et al., 2022). A2Perf uses the CodeCarbon library (Initiative, 2021) to track metrics during training, such as energy usage, power draw, RAM consumption, and wall-clock time. Energy and power usage inform the user about the sustainability and costs associated with training the agent, which is particularly important in power-constrained environments or when planning for long-term, continuous training (Parisi et al., 2019). RAM consumption metrics help in understanding the memory efficiency of the training process, as high RAM consumption may limit the settings where the agent can be trained or require costly hardware upgrades (Li et al., 2023). During the inference phase, A2Perf records power draw, RAM consumption, and average inference time (see Appendix A.4 for metric reporting guidelines and Appendix F for our full experimental setup).

2.3. Reliability

Reliability signifies safety, accountability, reproducibility, stability, and trustworthiness (Chan et al., 2019; Roszel et al., 2021). A2Perf uses the statistical methods proposed by Chan et al. (2019) to measure the reliability of autonomous agents during training and inference. For a detailed description of each metric and their calculation, please refer to the work by Chan et al. (2019).

2.4. Application Performance

Application performance is measured using task performance and generalization. Task performance is the agent's mean returns when rolled out for 100 episodes on the task it was trained for. Since autonomous agents deployed in real-world settings must often handle scenarios that differ from their exact training distribution, measuring generalization to tasks outside this distribution is crucial. Generalization is computed as the sum of mean returns for all tasks, including the task the agent was trained to perform.

3. Evaluation

Our evaluation aims to answer three key questions. (**Q1**) How can data cost metrics be used in practice to compare methods that use offline data to those that do not? (**Q2**) How do system performance metrics inform training and deployment feasibility? (**Q3**) Can reliability metrics reveal tradeoffs between different agents that are not captured by raw task performance?

For all domains and tasks, results are averaged over ten random seeds to ensure robustness and reproducibility. See Appendix E for more experimental results.

3.1. Q1: Comparing Across Algorithm Types with Data Cost

A2Perf provides datasets generated with agents of varying expertise (Section 2.1), along with their associated training sample costs. This enables the comparison of agents by considering both task performance and the cost of acquiring training data, which can vary significantly across different approaches like IL and RL.

Our experiments in the chip floorplanning domain reveal important insights about the true costs of different approaches. While BC's performance is competitive with DDQN and PPO (Appendix Table 4), the training sample cost – measured as the average energy consumed to train an agent that generates the data – was 48.28 kWh. In contrast, on-

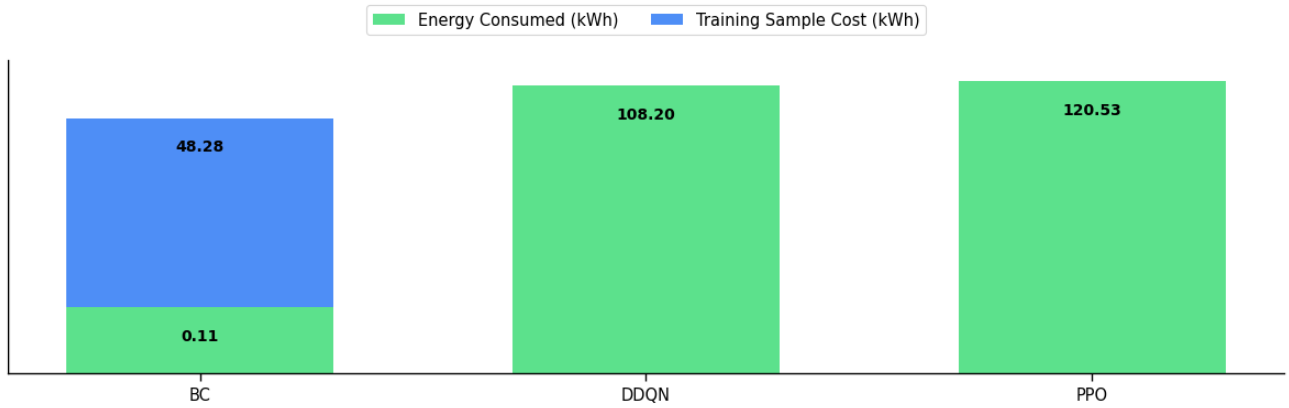

Figure 1: Comparison of energy consumption and training-sample cost for BC, DDQN, and PPO on the Ariane Netlist task, enabled by A2Perf. **Note:** The plot is not to scale for visibility of smaller values. Online methods (DDQN and PPO) have no training-sample cost because they are initialized without pre-collected data. BC's energy consumption (0.11 kWh) is significantly smaller than its training-sample cost (48.28 kWh), which represents the energy used to generate the training data.

line methods like DDQN and PPO learn purely through environment interaction without requiring any pre-collected datasets, resulting in a training sample cost of zero.

The data cost metric allows researchers to combine the training sample cost with the energy consumed during training for a more comprehensive comparison. This approach provides a total energy cost that can be directly compared across offline, online, or hybrid methods. For example, offline training of a BC agent for the Ariane netlist consumed only 0.11 kWh. Therefore, the total energy cost for a BC agent would be 48.39 kWh (48.28 kWh for generating the offline data + 0.11 kWh for offline training).

When comparing total energy costs, we find that despite requiring pre-collected data, BC's total energy cost (48.39 kWh) is still lower than the energy consumed by online methods like DDQN and PPO, which amounted to 108.20 kWh and 120.53 kWh, respectively (Figure 1). For hybrid methods that use both offline data and online environment interactions, the total energy cost would similarly be calculated by adding the training sample cost for the offline data to the energy consumed during the online training phase.

3.2. Q2: System Performance for Training and Deployment Feasibility

Our experiments in the web navigation domain highlight the importance of considering hardware constraints and performance requirements of autonomous agents. During training, PPO agents had a peak RAM usage of 2.3 ± 0.14 TB (Appendix Table 10). This high memory footprint can be attributed to the need for distributed experiments running hundreds of Google Chrome processes and storing batches

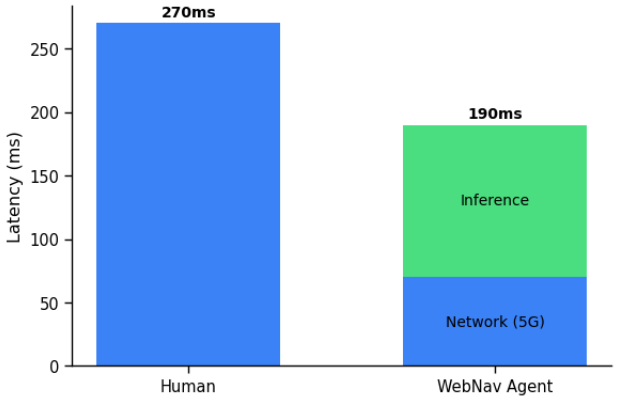

Figure 2: Latency of web-navigation agents compared with typical human reaction time. Even when served from the cloud, the agents respond quickly enough for real-time form-filling and similar interactive tasks.

of data, which involves tokenizing the entire DOM[5] tree of HTML elements on each web page. Such memory demands can limit the accessibility of training agents, as not all researchers may have access to the necessary hardware resources. To put this into perspective, training a variant of the GPT-3 language model with approximately 72 billion parameters would require a similar amount of memory, assuming each parameter is stored as a 32-bit floating-point number (Brown et al., 2020).

However, the resource usage of these agents becomes more manageable for deployment. The 120 ms inference time, when combined with the median round-trip latency of ∼68

[5] https://en.wikipedia.org/wiki/Document_Object_Model

ms for a 5G network (Schafhalter et al., 2023), results in a total latency of $\sim$200 ms. This combined latency is still faster than the average human reaction time of $\sim$273 ms[6], enabling real-time responsiveness during web navigation tasks (Figure 2). Furthermore, the peak RAM usage of 2.19 ± 0.09 GB (Appendix Table 10) indicates the feasibility of deploying trained agents directly on consumer-grade devices, such as smartphones, though the inference time may be slower on-device.

3.3. Q3: Finding Tradeoffs with Reliability Metrics

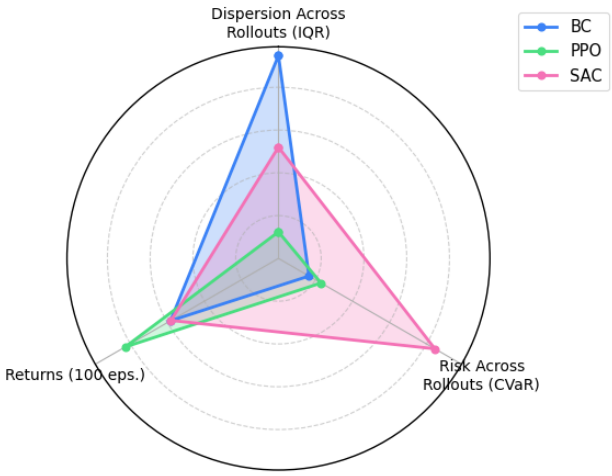

Figure 3: Reliability metrics for quadruped locomotion algorithms during inference on the **dog-pace** task. SAC shows markedly better reliability than PPO, achieving a $3.7\times$ lower worst-case return and a $1.8\times$ lower dispersion across rollouts.

In analyzing the "Dog Pace" task of QuadrupedLocomotion-v0 (Appendix Table 7), we observe overlapping error bars on the returns for PPO and SAC. To better understand their tradeoffs, we use the reliability metrics. PPO provides a 2x reduction in both short-term and long-term risks compared to SAC, making PPO more stable. This stability potentially makes PPO a safer option for training quadrupeds in the real world, where less sporadic behavior is needed. Conversely, SAC performs 3.7x better than PPO in the worst-case rollouts on average and demonstrates a 1.8x improvement in dispersion across rollouts, indicating more consistent gaits during deployment – essential from a safety perspective (Figure 3).

4. Conclusion

We need more holistic metrics and representative benchmarks to measure progress. To this end, we introduced A2Perf, a benchmarking suite that can be used for evaluating autonomous agents on challenging tasks from domains such as computer chip floorplanning, web navigation, and quadruped locomotion. A2Perf provides a standardized set of metrics across data cost, application performance, system resource efficiency, and reliability, enabling a comprehensive comparison of different algorithms. Our evaluations demonstrate A2Perf's effectiveness in identifying the strengths and weaknesses of various approaches to developing autonomous agents. We encourage the community to contribute new domains, tasks, and algorithms to A2Perf, making it an even more comprehensive platform for benchmarking autonomous agents in real-world-inspired settings.

[6] `https://humanbenchmark.com/tests/reactiontime/statistics`

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: a system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bonnet, C., Luo, D., Byrne, D., Surana, S., Abramowitz, S., Duckworth, P., Coyette, V., Midgley, L. I., Tegegn, E., Kalloniatis, T., et al. Jumanji: a diverse suite of scalable reinforcement learning environments in jax. *arXiv preprint arXiv:2306.09884*, 2023.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs. `http://github.com/google/jax`, 2018. Version 0.3.13.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.,

Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Chan, S. C., Fishman, S., Canny, J., Korattikara, A., and Guadarrama, S. Measuring the reliability of reinforcement learning algorithms. *arXiv preprint arXiv:1912.05663*, 2019.

Chevalier-Boisvert, M., Dai, B., Towers, M., de Lazcano, R., Willems, L., Lahlou, S., Pal, S., Castro, P. S., and Terry, J. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 1282–1289. PMLR, 2019.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Coleman, C. A., Narayanan, D., Kang, D., Zhao, T., Zhang, J., Nardi, L., Bailis, P. D., Olukotun, K., Ré, C., and Zaharia, M. A. Dawnbench : An end-to-end deep learning benchmark and competition. 2017. URL https://api.semanticscholar.org/CorpusID:3758333.

Coumans, E. Motion imitation, 2023. URL https://github.com/erwincoumans/motion_imitation.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Frey, N. C., Li, B., McDonald, J., Zhao, D., Jones, M., Bestor, D., Tiwari, D., Gadepally, V., and Samsi, S. Benchmarking resource usage for efficient distributed deep learning, 2022.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

Greaves, J., Candido, S., Dumoulin, V., Goroshin, R., Ponda, S. S., Bellemare, M. G., and Castro, P. S. Balloon Learning Environment, 12 2021. URL https://github.com/google/balloon-learning-environment.

Guadarrama, S., Korattikara, A., Ramirez, O., Castro, P., Holly, E., Fishman, S., Wang, K., Gonina, E., Wu, N., Kokiopoulou, E., Sbaiz, L., Smith, J., Bartók, G., Berent, J., Harris, C., Vanhoucke, V., and Brevdo, E. TF-Agents: A library for reinforcement learning in tensorflow. https://github.com/tensorflow/agents, 2018. URL https://github.com/tensorflow/agents. [Online; accessed 25-June-2019].

Guadarrama, S., Yue, S., Boyd, T., Jiang, J. W., Songhori, E., Tam, T., and Mirhoseini, A. Circuit Training: An open-source framework for generating chip floor plans with distributed deep reinforcement learning. https://github.com/google_research/circuit_training, 2021. URL https://github.com/google_research/circuit_training. [Online; accessed 21-December-2021].

Gulcehre, C., Wang, Z., Novikov, A., Paine, T., Gómez, S., Zolna, K., Agarwal, R., Merel, J. S., Mankowitz, D. J., Paduraru, C., et al. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:7248–7259, 2020.

Gur, I., Rueckert, U., Faust, A., and Hakkani-Tur, D. Learning to navigate the web. *arXiv preprint arXiv:1812.09195*, 2018.

Gur, I., Jaques, N., Miao, Y., Choi, J., Tiwari, M., Lee, H., and Faust, A. Environment generation for zero-shot compositional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:4157–4169, 2021.

Gur, I., Nachum, O., Miao, Y., Safdari, M., Huang, A., Chowdhery, A., Narang, S., Fiedel, N., and Faust, A. Understanding html with large language models. *arXiv preprint arXiv:2210.03945*, 2022.

Guss, W. H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M., and Salakhutdinov, R. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Initiative, M. Codecarbon. https://github.com/mlco2/codecarbon, 2021. Accessed: June 1, 2023.

Ji, J., Zhang, B., Zhou, J., Pan, X., Huang, W., Sun, R., Geng, Y., Zhong, Y., Dai, J., and Yang, Y. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36, 2023.

Kempka, M., Wydmuch, M., Runc, G., Toczek, J., and Jaśkowski, W. Vizdoom: A doom-based ai research platform for visual reinforcement learning, 2016.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Krishnan, S., Lam, M., Chitlangia, S., Wan, Z., Barth-Maron, G., Faust, A., and Reddi, V. J. Quarl: Quantization for fast and environmentally sustainable reinforcement learning, 2022.

Lee, Y., Hu, E. S., and Lim, J. J. Ikea furniture assembly environment for long-horizon complex manipulation tasks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6343–6349. IEEE, 2021.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, S., Tian, C., Tam, K., Ma, R., and Li, L. Breaking on-device training memory wall: A systematic survey, 2023.

Liu, E. Z., Guu, K., Pasupat, P., Shi, T., and Liang, P. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1802.08802.

Liu, Z., Guo, Z., Lin, H., Yao, Y., Zhu, J., Cen, Z., Hu, H., Yu, W., Zhang, T., Tan, J., et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023.

Mandlekar, A., Xu, D., Martín-Martín, R., Savarese, S., and Fei-Fei, L. Learning to generalize across long-horizon tasks from human demonstrations. *arXiv preprint arXiv:2003.06085*, 2020.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Bae, S., et al. Chip placement with deep reinforcement learning. *arXiv preprint arXiv:2004.10746*, 2020.

Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Nazi, A., et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Park, S., Frans, K., Eysenbach, B., and Levine, S. Ogbench: Benchmarking offline goal-conditioned rl. *arXiv preprint arXiv:2410.20092*, 2024.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Patterson, D. How we're minimizing ai's carbon footprint. https://blog.google/technology/ai/minimizing-carbon-footprint/. (Accessed on 06/05/2024).

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. Carbon emissions and large neural network training, 2021.

Peng, X. B., Coumans, E., Zhang, T., Lee, T.-W., Tan, J., and Levine, S. Learning agile robotic locomotion skills by imitating animals. *arXiv preprint arXiv:2004.00784*, 2020.

Qin, R.-J., Zhang, X., Gao, S., Chen, X.-H., Li, Z., Zhang, W., and Yu, Y. Neorl: A near real-world benchmark for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24753–24765, 2022.

Rafailov, R., Hatch, K., Singh, A., Smith, L., Kumar, A., Kostrikov, I., Hansen-Estruch, P., Kolev, V., Ball, P., Wu, J., et al. D5rl: Diverse datasets for data-driven deep reinforcement learning. *arXiv preprint arXiv:2408.08441*, 2024.

Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.-J., Anderson, B., Breughe, M., Charlebois, M., Chou, W., et al. Mlperf inference benchmark. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pp. 446–459. IEEE, 2020.

Roszel, M., Norvill, R., Hilger, J., and State, R. Know your model (kym): Increasing trust in ai and machine learning. *arXiv preprint arXiv:2106.11036*, 2021.

Schafhalter, P., Kalra, S., Xu, L., Gonzalez, J. E., and Stoica, I. Leveraging cloud computing to make autonomous vehicles safer. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5559–5566. IEEE, 2023.

Shi, T., Karpathy, A., Fan, L., Hernandez, J., and Liang, P. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. Deepmind control suite, 2018.

Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G. d., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. Gymnasium, March 2023. URL `https://zenodo.org/record/8127025`.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. Mastering atari games with limited data. *Advances in Neural Information Processing Systems*, 34:25476–25488, 2021.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Appendix

Table of Contents

1 Introduction **1**

2 Metrics for Real-World Evaluation **2**
- 2.1 Data Cost . . . 2
- 2.2 System Performance . . . 3
- 2.3 Reliability . . . 3
- 2.4 Application Performance . . . 3

3 Evaluation **3**
- 3.1 **Q1**: Comparing Across Algorithm Types with Data Cost . . . 3
- 3.2 **Q2**: System Performance for Training and Deployment Feasibility . . . 4
- 3.3 **Q3**: Finding Tradeoffs with Reliability Metrics . . . 4

4 Conclusion **5**

Appendix **9**

A. Metrics

A.1. Overview

	Data Cost	System	Reliability	Application
Training	Training Sample Cost	Energy Power RAM Usage Wall-Clock Time	Dispersion (Runs) Dispersion (Time) Long-Term Risk (Time) Risk (Runs) Short-Term Risk (Time)	Episodic Returns Generalization Returns
Inference	N/A	Inference Time Power RAM Usage	Dispersion (Rollouts) Risk (Rollouts)	N/A

Table 2: A2Perf assesses four categories—data cost, system performance, reliability, and application performance—during training and inference. These metrics provide a comprehensive evaluation of autonomous agents. See Section 2 for detailed descriptions of the metric categories. Data Cost is marked as "N/A" at inference time since pre-existing data and demonstrations are only used during training. Application metrics are marked as "N/A" during inference since performance and generalization are evaluated based on the complete training process.

A.2. Applying Metrics to Guide Real-World Algorithm Selection

Our experimental evaluation across three domains demonstrates how different metrics in A2Perf reveal complementary aspects of autonomous agent performance that are essential for real-world deployment decisions. While task performance provides a necessary baseline for comparison, the additional dimensions of data cost, system performance, and reliability metrics offer crucial insights for practitioners making algorithm selection decisions.

The relative importance of different metrics varies significantly by domain, reflecting different priorities in real-world applications. For chip floorplanning, reliability metrics revealed PPO's advantage in providing consistent initial layouts—a

property not evident from task performance alone but critical for human designers who need predictable starting points. In web navigation, system performance metrics demonstrated that while training requires substantial compute resources, inference can be performed efficiently enough for real-time web interaction. For quadruped locomotion, reliability metrics exposed a fundamental tradeoff between PPO's training stability and SAC's deployment stability that would be entirely missed by examining only average returns.

These findings illustrate A2Perf's value as a comprehensive evaluation framework that enables informed algorithm selection based on application-specific priorities. Researchers developing new autonomous agent algorithms should consider which metrics matter most for their target domains: data-intensive applications may prioritize training sample cost, resource-constrained deployments might emphasize inference efficiency, safety-critical systems would focus on reliability metrics, and applications requiring adaptability would value generalization performance. By providing this multidimensional perspective, A2Perf helps bridge the gap between algorithm development and successful real-world deployment of autonomous agents.

A.3. Using A2Perf Metrics in Practice

The metrics provided by A2Perf across data cost, application performance, system performance, and reliability offer a holistic view of an agent's performance. However, the relative importance of these metrics can vary significantly depending on the specific application domain. For instance, in resource-constrained environments, system performance metrics may be critical, while in safety-critical applications, reliability metrics might take precedence. In Section 3, we demonstrate how these metrics can be applied and interpreted in the context of our three benchmark domains: computer chip floorplanning, web navigation, and quadruped locomotion.

A.4. Community Benchmarking with A2Perf

While system performance metrics like energy usage, inference time, and memory consumption can vary significantly across different hardware platforms and software implementations, these measurements become meaningful when properly contextualized. To facilitate fair and useful comparisons, A2Perf will include a community leaderboard where researchers must report:

- **Hardware Configuration:**
  - CPU model
  - GPU model
- **Software Environment:**
  - Deep learning framework (e.g., PyTorch (Paszke et al., 2019), Tensorflow (Abadi et al., 2016), Jax (Bradbury et al., 2018), etc.)
  - Python Version
  - Operating System
- **Metric Results:**
  - Data Cost
  - System Performance
  - Reliability
  - Application Performance
- **Experimental Details:**
  - Number of random seeds used
  - All Hyperparameter Settings

To facilitate this standardized reporting and obtain the metric results above, researchers can leverage A2Perf's easy-to-use, open-source codebase. The codebase includes detailed tutorials, examples, and docker containers that simplify the evaluation process. Its modular implementation allows users to integrate their own custom algorithms without needing to modify the benchmarking code. As a result of this flexible design, A2Perf seamlessly integrates with existing tools for experiment

tracking such as TensorBoard[7], WandB[8], or MLFlow[9], allowing researchers to continue using their preferred development and visualization tools while still benefiting from A2Perf's standardized evaluation framework.

By standardizing the reporting of system configurations, researchers can meaningfully compare results across similar hardware and software setups, providing insights into how different agents perform under comparable conditions. The community leaderboard also enables understanding of performance scaling across different platforms, from resource-constrained environments to high-performance systems. Furthermore, practitioners can use this information to make informed decisions about deployment requirements and track optimization progress for specific hardware targets.

For example, researchers deploying a quadruped with a specific compute stack could filter the leaderboard entries to find results from comparable system configurations. As the community contributes results to A2Perf, this repository of performance data will expand across many computing environments, providing comprehensive coverage of different configurations.

B. Related Work

Benchmarking Autonomous Agents Table 1 offers a comparison between A2Perf and existing benchmarks, highlighting the unique contributions of our proposed benchmarking suite. Existing benchmarks for autonomous agents, such as those introduced by Brockman et al. (2016); Bellemare et al. (2013); Tassa et al. (2018), provide diverse environments for testing various algorithms. However, these benchmarks often focus on specific types of learning algorithms or on evaluating particular desirable qualities in autonomous agents. For example, Fu et al. (2020) and Gulcehre et al. (2020) evaluate offline reinforcement learning (Levine et al., 2020), while Yu et al. (2020) focuses on meta-reinforcement learning (Wang et al., 2016). Similarly, Ye et al. (2021) tests sample efficiency, Guss et al. (2019) challenges agents on long-horizon tasks, and Cobbe et al. (2019) evaluates generalization ability. While these benchmarks provide insights, they do not fully capture the challenges faced by autonomous agents in real-world applications (Dulac-Arnold et al., 2021). Environments, benchmarks, and datasets have been made to foster the development of autonomous agents in real-world scenarios, such as aerial balloon navigation (Greaves et al., 2021), autonomous driving (Sun et al., 2020), website navigation (Gur et al., 2021), and furniture assembly (Lee et al., 2021). Yet, these initiatives are often domain-specific and lack the comprehensive scope needed to evaluate agents across a wide range of real-world challenges as outlined by prior work (Dulac-Arnold et al., 2021), which forms the basis for our work.

Consequently, there remains a need for more benchmarking suites that encompass a diverse set of tasks and environments, reflecting the complexity and variety of problems encountered in real-world applications. Among recent benchmarks, NeoRL (Qin et al., 2022) provides realistic environments for stock trading, utility management and industrial control, while OGBench (Park et al., 2024) emphasizes realistic tasks in the offline, goal-conditioned setting. Jumanji (Bonnet et al., 2023) focuses on providing fast, JAX-accelerated (Bradbury et al., 2018) implementations of combinatorial optimization tasks inspired by industry applications.

A2Perf differentiates itself by incorporating different real-world domains such as web navigation and computer chip floorplanning, while also including system performance, data cost, and reliability metrics in a unified package. This comprehensive approach allows for a more holistic evaluation of autonomous agents across diverse, practically relevant tasks and crucial deployment considerations.

Benchmarking System Performance In addition to evaluating task-specific performance metrics, analyzing the end-to-end performance cost and examining the hardware resources required to apply learning algorithms on specific environments has gained significant attention (Wu et al., 2022; Patterson). Benchmarks such as MLPerf (Reddi et al., 2020) and DAWNBench (Coleman et al., 2017) have been developed to assess various aspects of commercial deep learning workloads across training and inference, considering a diverse class of systems. Furthermore, recent studies have investigated the environmental impact of deep learning by quantifying the carbon footprint associated with training and inference using large neural network models (Patterson et al., 2021). This line of research has also extended to autonomous agents, with works like QuaRL demonstrating reduced energy consumption and emissions through lower-precision distributed training (Krishnan et al., 2022). Despite these efforts, there remains a need for evaluating the system performance and energy

[7] https://www.tensorflow.org/tensorboard
[8] https://wandb.ai
[9] https://mlflow.org

Real-World Challenges	Chip Floorplanning	Web Navigation	Quadruped Locomotion
(RW1)* Training offline from fixed logs.	✓	✓	✓
(RW2) Learning on the real system from limited samples.	✗	✗	✓
(RW3) High-dimensional and continuous state and action spaces.	✓	✗	✓
(RW4) Safety constraints.	✗	✓	✓
(RW5) Tasks are partially observable, non-stationary or stochastic.	✗	✗	✓
(RW6) Unspecified, multi-objective or risk sensitive reward functions.	✓	✓	✓
(RW7) Need for explainable policies.	✗	✓	✗
(RW8) Real-time inference at the control frequency of the system.	✗	✓	✓
(RW9) Delays in actuators, sensors or rewards.	✗	✓	✓

Table 3: Real-World Challenges proposed by Dulac-Arnold et al. (Dulac-Arnold et al., 2021). Checkmarks (✓) indicate challenges commonly encountered in the general domain area, while (✗) denotes challenges less frequently encountered. The challenge marked with an asterisk (*), RW1, applies to all A2Perf domains, as learning from offline data is possible for all environments. Each broad challenge is encountered in at least one of the A2Perf domain areas, highlighting the relevance of the selected domains to current real-world reinforcement learning problems.

consumption of autonomous agents to provide valuable insights into their practical feasibility and sustainability.

Reliability Metrics for Reinforcement Learning Reliability is a concern in reinforcement learning (RL), as current metrics often rely on point estimates of aggregate performance, which fail to capture the true performance of algorithms and make it challenging to draw conclusions about the state-of-the-art (Agarwal et al., 2021; Henderson et al., 2018; Colas et al., 2018). The increasing complexity of benchmarking tasks has made it infeasible to run hundreds of training runs, necessitating the development of tools to evaluate reliability based on a limited number of runs (Agarwal et al., 2021). For real-world deployments, reliability is essential to ensure that RL algorithms perform consistently and robustly across different conditions and environments. To assess reliability, it is essential to consider metrics across three axes of variability: time (within a training run), runs (across random seeds), and rollouts of a fixed policy (Chan et al., 2019). By incorporating reliability metrics into A2Perf, we will be able to better assess the robustness and consistency of RL algorithms.

C. A2Perf Domains

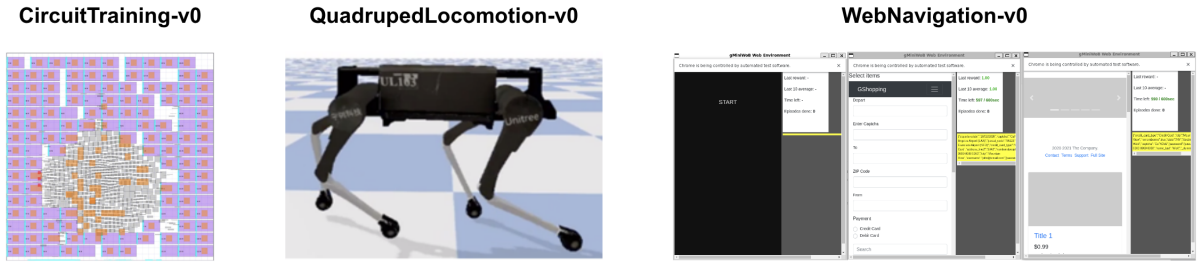

Figure 4: The three domains included in A2Perf: computer chip floorplanning for optimizing integrated circuit layouts, web navigation for automated form filling and website interaction, and quadruped locomotion for robotic control. These specific domains were selected based on their demonstrated transfer from simulation to real-world applications.

Our guiding question when selecting domains for A2Perf was "how can we choose domains that reflect real-world applications of autonomous agents?" To identify suitable domains, we conducted interviews with industry practitioners to understand where autonomous agents are currently deployed and where they show future promise. This process led us to three application areas with significant industrial relevance: computer chip floorplanning, quadruped locomotion, and

website navigation.

From these industrially relevant application areas, we specifically selected domains with demonstrated simulation-to-reality transfer. This selection criterion enables researchers without access to specialized hardware (like robots or chip fabrication facilities) to make meaningful contributions using simulated environments. The circuit training domain was used in creating an iteration of Google's Tensor Processing Unit (TPU) (Mirhoseini et al., 2021). The quadruped locomotion domain has been shown to transfer successfully to real Unitree Laikago robots (Peng et al., 2020). The web navigation domain is derived from MiniWob (Shi et al., 2017), MiniWob++ (Liu et al., 2018), and gMiniWob (Gur et al., 2021), and operates in an actual Google Chrome browser, mirroring real-life web interactions. Additionally, (Gur et al., 2018) showed that policies trained in MiniWob++ transfer to real-life web pages for task completion.

By focusing on domains with demonstrated real-world applicability, progress made within the A2Perf benchmark can directly contribute to improving the performance of downstream real-world (RW) tasks. We specify how each domain aligns with the real-world challenges presented by Dulac-Arnold et al. (2021) (Table 3), and denote which of A2Perf's metric categories are important for each domain.

C.1. Circuit Training (RW1, RW3, RW6)

Chip floorplanning involves creating a physical layout for a microprocessor, a task that has resisted automation for decades and requires months of human engineering effort. To address this challenge, Google has made Circuit Training available as an open-source framework that uses RL to generate chip floorplans (Guadarrama et al., 2021). In this domain, an agent places macros (reusable blocks of circuitry) onto the chip canvas, with the objective of optimizing wirelength, congestion, and density. Even though the state and action spaces are discrete, the number of states and actions increases combinatorially with the number of nodes and cells on the chip (RW3). As an illustration, Mirhoseini et al. (2021) calculate that placing 1,000 clusters of nodes on a grid with 1,000 cells results in a state space on the order of $10^{2,500}$, which is vastly larger than the state space of Go at 10^{360}. Chip design also involves optimizing for multiple objectives, such as maximizing clock frequency, reducing power consumption, and minimizing chip area (RW6). During training, these objectives are approximated using proxy metrics. However, evaluating the true objectives requires time-consuming simulations with industry-grade placement tools [10]. If the results are unsatisfactory, the proxy metrics must be adjusted, and the agents must be retrained, leading to a costly iterative and resource-intensive process.

Important Metric Categories For Circuit Training agents, the following metric categories are most critical for real-world use:

- **Task Performance**: Circuit Training agents must generate high-quality macro placements by minimizing wirelength, congestion, and density of the chip.
- **Inference Reliability**: Chip designers use these agents to generate initial macro placements, then manually refine them. Agents must produce consistent macro placements across multiple rollouts. Inconsistent placements would force designers to repeatedly roll out the same policy to try achieving favored initial placements.
- **Inference System Performance**: Fast inference time is crucial to enable interactive use by human designers. Designers need to quickly evaluate and refine different placement options.
- **Generalization**: The ability to handle new circuit architectures without retraining is vital, as new circuits are frequently created. Strong generalization performance reduces the need to train separate agents for each new netlist.
- **Data Cost**: Many circuit netlists are proprietary, and generating high-quality macro placements requires significant human effort. Understanding data collection costs helps evaluate the practicality of different learning approaches.

C.2. Web Navigation (RW1, RW4, RW6, RW7, RW8, RW9)

Software tools exist to automate browser tasks[11], but due to the varied formatting of websites, hand-crafted algorithms are not a viable solution for general web navigation. Researchers have begun applying learning algorithms to design agents that can understand web pages (Gur et al., 2022) and automatically navigate through them to fill out forms (Gur et al., 2021;

[10] For example, Cadence Innovus and Synopsys IC Compiler

[11] Selenium, used in A2Perf, is a popular browser automation tool.

2018). In A2Perf, we use gMiniWob (Gur et al., 2021) to create mock websites that act as environments for the agent. See Appendix K for details about the website generation process and agent interaction. To achieve maximum rewards, the agent must avoid malicious links and advertisement banners (RW4) while correctly filling out all fields in web forms. The combination of these constraints create a multi-objective reward function (RW6). The explainability of an agent's decision-making is also important, particularly when agents handle sensitive tasks such as online shopping or investing (RW7). Finally, agents must be robust to the system challenges of real-time inference, such as inference speed and network delays (RW8, RW9).

Important Metric Categories For web navigation agents, the following metric categories are most critical for real-world use:

- **Task Performance**: Agents must accurately complete web forms and navigate sites correctly.
- **Inference System Performance**: Agents need to operate at speeds comparable to human web browsing to provide a seamless user experience. This includes both inference time and resource usage on consumer devices.
- **Inference Reliability**: Reliability is crucial for safety, as unreliable agents might occasionally click on malicious links or advertisements. Even rare mistakes in web navigation can have serious consequences.
- **Generalization**: Websites vary greatly in design and structure. Agents must adapt to different layouts, styles, and interaction patterns without requiring retraining for each new site.
- **Training System Performance**: Web navigation training involves processing HTML pages and running multiple browser instances, creating significant computational demands.

C.3. Quadruped Locomotion (RW1, RW2, RW3, RW4, RW5, RW6, RW8, RW9)

In recent years, the robotics community has gradually shifted towards training autonomous agents for robotic control. A prominent example of this trend is seen in quadruped locomotion, where RL has become the dominant technique. We followed the work of Peng et al. (2020), in which a quadruped robot learns complex locomotion skills such as pacing, trotting, spinning, hop-turning, and side-stepping by imitating motion capture data from a real dog.

Given the physical dynamics involved in quadruped locomotion, research often necessitates learning directly from limited samples on the actual robot (RW2). Learning walking gaits also involves high-dimensional, continuous state and action spaces (RW3), as the robot needs to precisely control multiple joints and limbs to navigate complex environments. The agent must reason about complex dynamics, avoid unsafe falls (RW4), adapt gaits to various speeds and terrains (RW5), and operate in partially observable environments (RW5) where states like contact forces are not directly measurable. Optimizing robotic controllers is usually multi-objective (RW6), balancing competing objectives like locomotion speed, stability, satisfying safety constraints, and minimizing energy expenditure. Furthermore, real-time inference (RW8) and dealing with system delays (RW9) are critical for controlling robots, as slow computations or delays can negatively impact stability and performance.

Important Metric Categories For quadruped locomotion agents, the following metric categories are most critical for real-world use:

- **Task Performance**: Agents must accurately reproduce desired walking gaits, as poor imitation of natural movements can lead to inefficient or unstable locomotion.
- **Inference Reliability**: The agent must maintain smooth, stable motions without sudden movements or changes in behavior. Inconsistent movements could damage the robot or cause falls in real-world environments.
- **Inference System Performance**: Quadrupeds require real-time responsiveness from their onboard computers to maintain stability. Both inference speed and energy efficiency are crucial, as robots often operate with limited computing resources and battery power.
- **Generalization**: Robots must adapt to different terrains, slopes, and surface conditions without retraining. Strong generalization also helps robots handle variations in their own morphology due to wear or manufacturing differences.

Category	Metric Name	BC	DDQN	PPO
Ariane (Training)				
Data Cost	Training Sample Cost ↓	48.28	**0**	**0**
Application	Generalization (100 eps. [all tasks]) ↑	-2.18	-2.19	**-2.05**
	Returns (100 eps.) ↑	-1.10 ± 0.04	-1.13 ± 0.04	**-0.99 ± 7.25e-03**
Reliability	Dispersion Across Runs (IQR) ↓	N/A	**0.03 ± 0.03**	0.04 ± 0.02
	Dispersion Within Runs (IQR) ↓	N/A	0.02 ± 0.03	**4.77e-03 ± 4.92e-03**
	Long Term Risk (CVaR) ↓	N/A	1.20	**0.03**
	Risk Across Runs (CVaR) ↑	N/A	-1.17	**-1.03**
	Short Term Risk (CVaR) ↓	N/A	0.07	**0.01**
System	Energy Consumed (kWh) ↓	**0.11 ± 6.45e-04**	108.20 ± 4.29	120.53 ± 2.78
	GPU Power Usage (W) ↓	**211.35 ± 16.76**	585.98 ± 172.50	692.94 ± 120.08
	Mean RAM Usage (GB) ↓	**4.72 ± 0.53**	849.37 ± 64.85	834.05 ± 55.90
	Peak RAM Usage (GB) ↓	**5.25 ± 0.07**	889.56 ± 23.44	906.45 ± 68.01
	Wall Clock Time (Hours) ↓	**0.48 ± 2.61e-03**	21.94 ± 0.90	23.95 ± 0.54
Ariane (Inference)				
Reliability	Dispersion Across Rollouts (IQR) ↓	**0.01**	0.05	**0.01**
	Risk Across Rollouts (CVaR) ↑	-1.23	-1.25	**-1.01**
System	GPU Power Usage (W) ↓	136.91 ± 21.48	69.50 ± 4.60	**49.43 ± 30.29**
	Inference Time (ms) ↓	**10.0 ± 0.46**	20.0 ± 2.69	20.0 ± 2.68
	Mean RAM Usage (GB) ↓	2.19 ± 0.21	**2.15 ± 0.30**	2.51 ± 0.49
	Peak RAM Usage (GB) ↓	2.29 ± 0.01	**2.28 ± 0.13**	2.71 ± 0.62

Table 4: Metrics for the Ariane Netlist task of CircuitTraining-v0. All metrics are averaged over ten random seeds. We report mean and standard deviation for metrics where it is applicable. BC results are obtained by training on the entire `intermediate` dataset. Note that the training reliability metrics for BC are marked as "N/A" since BC does not perform online rollouts in the environment. The ↑ symbol indicates metrics where higher values are better, while ↓ indicates metrics where lower values are better. **Bold values** highlight the best performing algorithm for each metric.

D. Limitations and Future Work

A2Perf includes three domains that cover a diverse range of real-world applications and challenges, but there is room for expansion to a wider range of tasks. Thanks to A2Perf's integration with Gymnasium (Towers et al., 2023) (previously OpenAI Gym) and the implementation of baselines using TF-Agents (Guadarrama et al., 2018), adding new domains and baselines is straightforward, making it easy for researchers to contribute to the platform.

Future work could expand A2Perf to include multi-agent domains and tasks, reflecting real-world scenarios where autonomous agents interact with other agents and humans. Many real-world applications inherently involve multiple agents coordinating or competing: autonomous vehicles navigating in traffic, robot teams in warehouse settings, or trading agents in financial markets. Integrating multi-agent domains would require additional metrics to capture interaction dynamics, such as coordination efficiency, communication overhead, and emergent social behaviors. For example, extending the quadruped locomotion domain to include multiple robots collaboratively navigating complex terrain would test both individual control and collective coordination capabilities, potentially revealing new insights about algorithm robustness in social contexts.

Another area of future work is the addition of support for measuring system performance on custom hardware platforms. This would provide more precise insights into performance in target deployment environments, as current evaluations are conducted primarily on desktop and server machines. This extension is particularly important for edge computing applications such as robotics, where power constraints, thermal limitations, and specialized accelerators significantly influence real-world performance. By developing standardized benchmarking procedures for specific deployment platforms such as NVIDIA Jetson[12], Google Coral[13], or custom FPGA implementations, A2Perf could offer more accurate predictions

[12] `https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/`
[13] `https://coral.ai/`

of deployment performance. This would help bridge the gap between research prototypes and production systems by identifying specific optimization opportunities for the target hardware.

To further standardize evaluations in A2Perf, future work should address potential variations due to different hardware setups, Python versions, and code implementations. Even with our current efforts to ensure reproducibility, subtle differences in environment configurations can lead to meaningful performance variations. Creating a centralized evaluation server, similar to the approach taken by MLPerf (Reddi et al., 2020), could further standardize comparisons by running all submissions in identical environments. These enhancements would facilitate more accurate comparisons between different computing environments.

As an open-source platform, A2Perf is designed to evolve through community contributions. Researchers can extend the benchmark in multiple ways: by adding new domains through the standard Gymnasium interface, by implementing additional baseline algorithms, or by introducing domain-specific metrics that capture other aspects of real-world performance. The repository includes detailed contribution guidelines, templates, and documentation to facilitate these extensions. This collaborative approach ensures A2Perf remains relevant to emerging research challenges while expanding its coverage of real-world autonomous agent applications.

E. Additional Experiments

We present an extensive set of additional experiments that showcase A2Perf's capabilities in evaluating autonomous agents across various domains and tasks. The results encompass a wide range of metrics, including data cost, reliability, system performance, and application performance, providing a holistic view of the strengths and limitations of different algorithmic approaches.

The circuit training domain experiments (Appendix E.1) reveal interesting trade-offs between behavioral cloning, DDQN, and PPO in terms of data efficiency, computational requirements, and performance consistency. Moving to the quadruped locomotion domain (Appendix E.2), we observe how the reliability metrics shed light on the robustness and worst-case behavior of the agents during both training and inference phases. The web navigation domain (Appendix E.2) introduces an additional layer of complexity, with websites of varying difficulty levels. Here, the system performance metrics highlight the substantial computational demands, particularly in terms of memory usage, associated with training web navigation agents. To further facilitate a clear and intuitive comparison of the algorithms' performance across all domains and tasks, we have included graphical visualizations (Appendix E.4) that summarize the key metrics along different evaluation dimensions.

These experiments show A2Perf's versatility in providing a comprehensive and nuanced evaluation of autonomous agents operating in diverse and realistic settings. By considering multiple performance aspects and presenting the results in both tabular and graphical formats, A2Perf enables researchers and practitioners to gain valuable insights into the behavior and limitations of different algorithmic choices, ultimately guiding the development of more robust and efficient autonomous agents.

E.1. Circuit Training

This section shows the full set of metrics for the toy macro standard cell and Ariane netlists in the circuit training domain. The results highlight the differences in data cost, reliability, system performance, and application performance between behavioral cloning (BC), DDQN, and PPO.

E.2. Quadruped Locomotion

This section reports the metrics for the dog pace, trot, and spin gaits in the quadruped locomotion domain. The reliability metrics provide insights into the stability and worst-case performance of the algorithms during training and inference.

E.3. Web Navigation

This section details the evaluation on websites of varying difficulty levels in the web navigation domain. The system performance metrics underscore the significant computational requirements, especially in terms of RAM usage, for training web navigation agents.

Category	Metric Name	BC	DDQN	PPO
Toy Macro Standard Cell (Training)				
Data Cost	Training Sample Cost (kWh)	4.44	0	0
Application	Generalization (100 eps. [all tasks])	-2.19	-2.20	-2.13
	Returns (100 eps.)	-0.97 $\pm$ 2.27 $\times 10^{-3}$	-1.05 $\pm$ 0.04	-0.97 $\pm$ 8.09 $\times 10^{-3}$
Reliability	Dispersion Across Runs (IQR)	N/A	0.01 $\pm$ 0.01	9.07e-03 $\pm$ 6.43 $\times 10^{-3}$
	Dispersion Within Runs (IQR)	N/A	8.80 $\times 10^{-3}$ $\pm$ 0.01	2.51 $\times 10^{-3}$ $\pm$ 3.61 $\times 10^{-3}$
	Long Term Risk (CVaR)	N/A	1.10	0.04
	Risk Across Runs (CVaR)	N/A	-1.08	-0.99
	Short Term Risk (CVaR)	N/A	0.03	9.89 $\times 10^{-3}$
System	Energy Consumed (kWh)	0.02 $\pm$ 1.97 $\times 10^{-4}$	5.55 $\pm$ 2.03	15.37 $\pm$ 3.79
	GPU Power Usage (W)	188.20 $\pm$ 21.98	448.00 $\pm$ 200.41	307.05 $\pm$ 69.75
	Peak RAM Usage (GB)	4.71 $\pm$ 0.02	525.99 $\pm$ 205.64	675.26 $\pm$ 45.30
	Wall Clock Time (Hours)	0.10 $\pm$ 1.36 $\times 10^{-3}$	0.29 $\pm$ 0.57	1.79 $\pm$ 2.16
Toy Macro Standard Cell (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	1.68 $\times 10^{-3}$	0.09	2.43 $\times 10^{-3}$
	Risk Across Rollouts (CVaR)	-0.97	-1.10	-0.99
System	GPU Power Usage (W)	104.97 $\pm$ 22.85	59.45 $\pm$ 1.43	58.97 $\pm$ 1.14
	Inference Time (ms)	8.93 $\pm$ 0.51	20 $\pm$ 2.69	20 $\pm$ 2.67
	Mean RAM Usage (GB)	1.92 $\pm$ 0.42	1.45 $\pm$ 0.48	1.99 $\pm$ 0.30
	Peak RAM Usage (GB)	2.14 $\pm$ 0.03	2.10 $\pm$ 0.05	2.16 $\pm$ 0.07

Table 5: Metrics for the "Toy Macro" netlist task of CircuitTraining-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

Category	Metric Name	BC	PPO	SAC
Dog Pace (Training)				
Data Cost	Training Sample Cost (kWh)	22.53	0	0
Application	Generalization (100 eps. [all tasks])	3.99	3.36	5.03
	Returns (100 eps.)	7.00 $\pm$ 4.68	9.94 $\pm$ 15.59	6.96 $\pm$ 6.72
Reliability	Dispersion Across Runs (IQR)	N/A	9.63 $\pm$ 7.27	3.61 $\pm$ 3.88
	Dispersion Within Runs (IQR)	N/A	2.22 $\pm$ 1.97	2.98 $\pm$ 3.64
	Long Term Risk (CVaR)	N/A	13.00	25.82
	Risk Across Runs (CVaR)	N/A	13.74	8.55
	Short Term Risk (CVaR)	N/A	5.81	10.19
System	Energy Consumed (kWh)	0.11 $\pm$ 0.02	32.46 $\pm$ 0.26	36.22 $\pm$ 2.33
	GPU Power Usage (W)	240.64 $\pm$ 5.41	280.12 $\pm$ 23.69	266.37 $\pm$ 9.54
	Mean RAM Usage (GB)	3.21 $\pm$ 0.24	532.93 $\pm$ 14.28	516.24 $\pm$ 75.03
	Peak RAM Usage (GB)	3.25 $\pm$ 0.01	534.26 $\pm$ 2.04	545.16 $\pm$ 0.50
	Wall Clock Time (Hours)	0.46 $\pm$ 0.07	18.73 $\pm$ 0.19	19.41 $\pm$ 2.74
Dog Pace (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	0.52	8.76	4.80
	Risk Across Rollouts (CVaR)	0.33	0.46	1.69
System	GPU Power Usage (W)	60.37 $\pm$ 1.78	59.11 $\pm$ 1.31	61.41 $\pm$ 1.96
	Inference Time (ms)	2.33 $\pm$ 0.54	2.56 $\pm$ 0.39	2.52 $\pm$ 0.74
	Mean RAM Usage (GB)	1.69 $\pm$ 0.31	1.81 $\pm$ 0.14	1.71 $\pm$ 0.30
	Peak RAM Usage (GB)	1.82 $\pm$ 0.03	1.84 $\pm$ 9.05e-03	1.85 $\pm$ 0.04

Table 7: Metrics for the "dog pace" gait of QuadrupedLocomotion-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

Dog Trot (Training)				
Category	**Metric Name**	**BC**	**PPO**	**SAC**
Data Cost	Training Sample Cost (kWh)	15.77	0	0
Application	Generalization (100 eps. [all tasks])	3.87	3.09	4.49
	Returns (100 eps.)	1.06 ± 0.26	1.49 ± 1.02	3.51 ± 2.88
Reliability	Dispersion Across Runs (IQR)	N/A	9.07 ± 4.93	0.85 ± 1.29
	Dispersion Within Runs (IQR)	N/A	0.82 ± 0.84	0.93 ± 1.11
	Long Term Risk (CVaR)	N/A	6.79	8.46
	Risk Across Runs (CVaR)	N/A	6.00	2.58
	Short Term Risk (CVaR)	N/A	2.41	3.20
System	Energy Consumed (kWh)	0.12 ± 0.02	16.82 ± 0.29	19.17 ± 0.64
	GPU Power Usage (W)	242.12 ± 7.53	277.71 ± 23.47	269.18 ± 10.12
	Mean RAM Usage (GB)	3.21 ± 0.25	535.00 ± 18.77	535.99 ± 29.49
	Peak RAM Usage (GB)	3.26 ± 0.01	536.47 ± 1.98	544.80 ± 4.39
	Wall Clock Time (Hours)	0.46 ± 0.06	18.57 ± 0.23	18.99 ± 6.78
Dog Trot (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	0.32	0.89	1.25
	Risk Across Rollouts (CVaR)	0.63	0.36	1.33
System	GPU Power Usage (W)	59.32 ± 1.08	58.91 ± 1.28	59.39 ± 1.23
	Inference Time (ms)	2.32 ± 0.49	2.55 ± 0.57	2.45 ± 0.35
	Mean RAM Usage (GB)	1.66 ± 0.33	1.76 ± 0.25	1.80 ± 0.17
	Peak RAM Usage (GB)	$1.82 \pm 8.77 \times 10^{-4}$	1.85 ± 0.02	1.85 ± 0.03

Table 8: Metrics for the "dog trot" gait of QuadrupedLocomotion-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

Category	Metric Name	BC	PPO	SAC
Dog Spin (Training)				
Data Cost	Training Sample Cost (kWh)	30.17	0	0
Application	Generalization (100 eps. [all tasks])	3.97	2.69	4.61
	Returns (100 eps.)	1.54 ± 0.42	3.82 ± 6.22	3.84 ± 1.46
Reliability	Dispersion Across Runs (IQR)	N/A	7.92 ± 4.60	0.74 ± 0.76
	Dispersion Within Runs (IQR)	N/A	1.00 ± 1.08	0.84 ± 1.26
	Long Term Risk (CVaR)	N/A	8.88	14.37
	Risk Across Runs (CVaR)	N/A	8.29	3.82
	Short Term Risk (CVaR)	N/A	3.09	2.99
System	Energy Consumed (kWh)	0.10 ± 0.04	17.42 ± 0.35	18.88 ± 0.59
	GPU Power Usage (W)	216.72 ± 68.63	278.38 ± 22.60	264.46 ± 9.49
	Mean RAM Usage (GB)	3.18 ± 0.26	534.56 ± 21.28	531.27 ± 55.64
	Peak RAM Usage (GB)	3.23 ± 0.08	536.10 ± 3.03	477.22 ± 172.63
	Wall Clock Time (Hours)	0.45 ± 0.08	17.13 ± 6.07	17.02 ± 9.05
Dog Spin (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	0.37	2.41	1.78
	Risk Across Rollouts (CVaR)	0.28	0.12	0.55
System	GPU Power Usage (W)	60.10 ± 1.14	59.70 ± 1.22	59.65 ± 1.73
	Inference Time (ms)	2.33 ± 0.66	2.45 ± 0.48	2.41 ± 0.22
	Mean RAM Usage (GB)	1.68 ± 0.32	1.79 ± 0.22	1.75 ± 0.26
	Peak RAM Usage (GB)	1.82 ± 0.03	1.85 ± 0.02	1.84 ± 0.02

Table 9: Metrics for the "dog spin" gait of QuadrupedLocomotion-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

Category	Metric Name	BC	DDQN	PPO
Difficulty 1, 1 Website (Training)				
Data Cost	Training Sample Cost (kWh)	14.15	0	0
Application	Generalization (100 eps. [all tasks])	-12.94	-11.15	-24.54
	Returns (100 eps.)	-3.57 $\pm$ 2.80	-7.55 $\pm$ 5.74	-13.45 $\pm$ 0.51
Reliability	Dispersion Across Runs (IQR)	N/A	0.73 $\pm$ 0.63	4.20 $\pm$ 1.45
	Dispersion Within Runs (IQR)	N/A	0.37 $\pm$ 0.68	0.57 $\pm$ 0.53
	Long Term Risk (CVaR)	N/A	9.32	12.12
	Risk Across Runs (CVaR)	N/A	-2.75	-13.11
	Short Term Risk (CVaR)	N/A	1.79	1.86
System	Energy Consumed (kWh)	$0.04 \pm 6.02 \times 10^{-4}$	29.56 $\pm$ 7.23	28.82 $\pm$ 1.19
	GPU Power Usage (W)	125.89 $\pm$ 2.53	265.09 $\pm$ 21.50	305.15 $\pm$ 34.41
	Mean RAM Usage (GB)	4.10 $\pm$ 0.33	1140.98 $\pm$ 580.55	1592.45 $\pm$ 388.64
	Peak RAM Usage (GB)	4.23 $\pm$ 0.04	1931.54 $\pm$ 242.31	2305.57 $\pm$ 135.48
	Wall Clock Time (Hours)	$0.31 \pm 4.91 \times 10^{-3}$	8.13 $\pm$ 5.17	10.50 $\pm$ 0.44
Difficulty 1, 1 Website (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	3.36	11.75	0.50
	Risk Across Rollouts (CVaR)	-10.65	-13.25	-13.75
System	GPU Power Usage (W)	108.61 $\pm$ 15.76	59.61 $\pm$ 1.41	60.26 $\pm$ 1.14
	Inference Time (ms)	3.07 $\pm$ 0.47	110 $\pm$ 9.93	120 $\pm$ 9.71
	Mean RAM Usage (GB)	1.97 $\pm$ 0.32	2.08 $\pm$ 0.20	2.12 $\pm$ 0.17
	Peak RAM Usage (GB)	2.11 $\pm$ 0.11	2.18 $\pm$ 0.11	2.19 $\pm$ 0.09

Table 10: Metrics for "difficulty 1, 1 website" task of WebNavigation-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

Category	Metric Name	BC	DDQN	PPO
Difficulty 1, 5 Websites (Training)				
Data Cost	Training Sample Cost (kWh)	13.66	0	0
Application	Generalization (100 eps. [all tasks])	-13.34	-11.03	-23.86
	Returns (100 eps.)	-4.87 $\pm$ 3.33	-3.43 $\pm$ 4.58	-12.37 $\pm$ 3.53
Reliability	Dispersion Across Runs (IQR)	N/A	0.43 $\pm$ 0.55	3.42 $\pm$ 1.08
	Dispersion Within Runs (IQR)	N/A	0.49 $\pm$ 0.97	0.75 $\pm$ 0.55
	Long Term Risk (CVaR)	N/A	11.27	11.70
	Risk Across Runs (CVaR)	N/A	-1.26	-12.60
	Short Term Risk (CVaR)	N/A	2.47	2.05
System	Energy Consumed (kWh)	$0.04 \pm 4.82 \times 10^{-4}$	31.59 $\pm$ 5.19	28.48 $\pm$ 1.22
	GPU Power Usage (W)	126.04 $\pm$ 4.03	265.81 $\pm$ 22.08	303.28 $\pm$ 34.99
	Mean RAM Usage (GB)	4.03 $\pm$ 0.34	1206.86 $\pm$ 466.37	1545.56 $\pm$ 427.22
	Peak RAM Usage (GB)	4.15 $\pm$ 0.11	1928.69 $\pm$ 209.62	2227.07 $\pm$ 210.77
	Wall Clock Time (Hours)	$0.30 \pm 3.71 \times 10^{-3}$	9.35 $\pm$ 4.70	10.45 $\pm$ 0.31
Difficulty 1, 5 Websites (Inference)				
Reliability	Dispersion Across Rollouts (IQR)	5.96	0.29	0.50
	Risk Across Rollouts (CVaR)	-11.36	-13.46	-13.75
System	GPU Power Usage (W)	108.13 $\pm$ 16.85	60.87 $\pm$ 5.78	60.17 $\pm$ 1.67
	Inference Time (ms)	3.04 $\pm$ 0.44	110 $\pm$ 9.83	120 $\pm$ 9.21
	Mean RAM Usage (GB)	1.97 $\pm$ 0.33	2.07 $\pm$ 0.32	2.12 $\pm$ 0.16
	Peak RAM Usage (GB)	2.12 $\pm$ 0.03	2.57 $\pm$ 0.86	2.19 $\pm$ 0.01

Table 11: Metrics for "difficulty 1, 5 websites" task of WebNavigation-v0. All metrics are averaged over ten random seeds. Note that the training reliability metrics for BC are marked as “N/A” since BC does not perform online rollouts in the environment.

E.4. Radar Plots for Easy Visual Comparison

These figures provide a graphical representation of the key metrics across all domains and tasks, enabling a visual comparison of the algorithms' performance along the different evaluation axes.

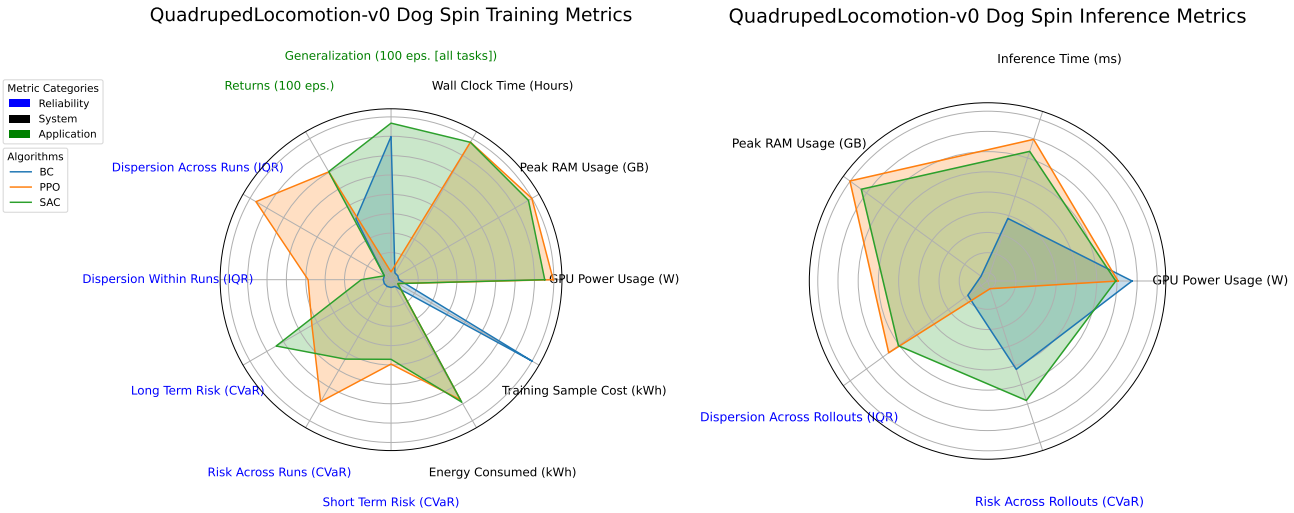

Figure 5: Graphical representation of metrics for the "dog spin" gait of QuadrupedLocomotion-v0

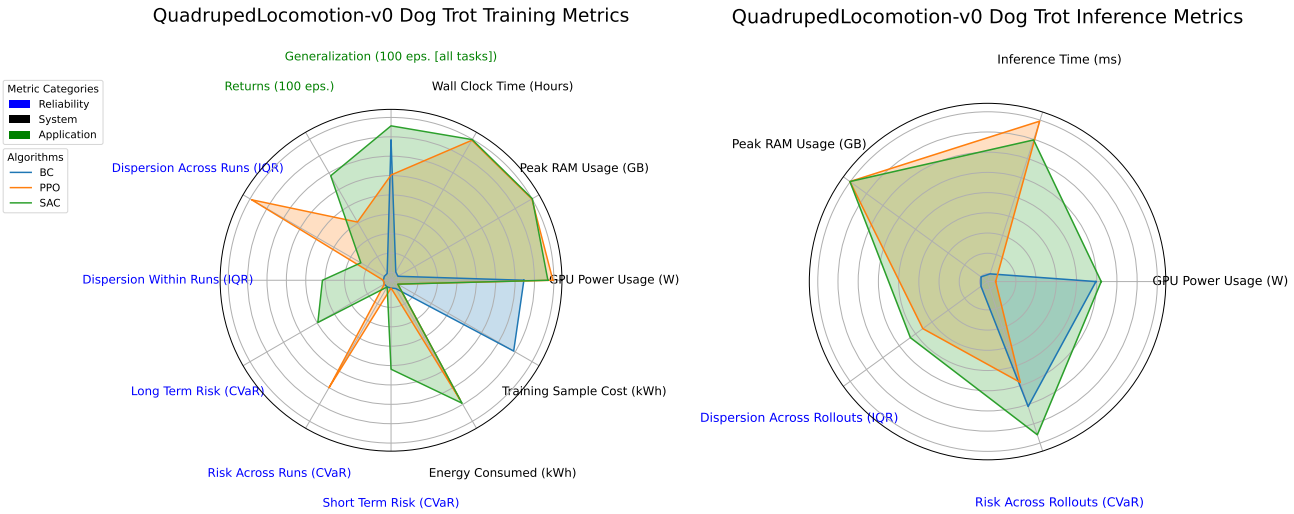

Figure 6: Graphical representation of metrics for the "dog trot" gait of QuadrupedLocomotion-v0

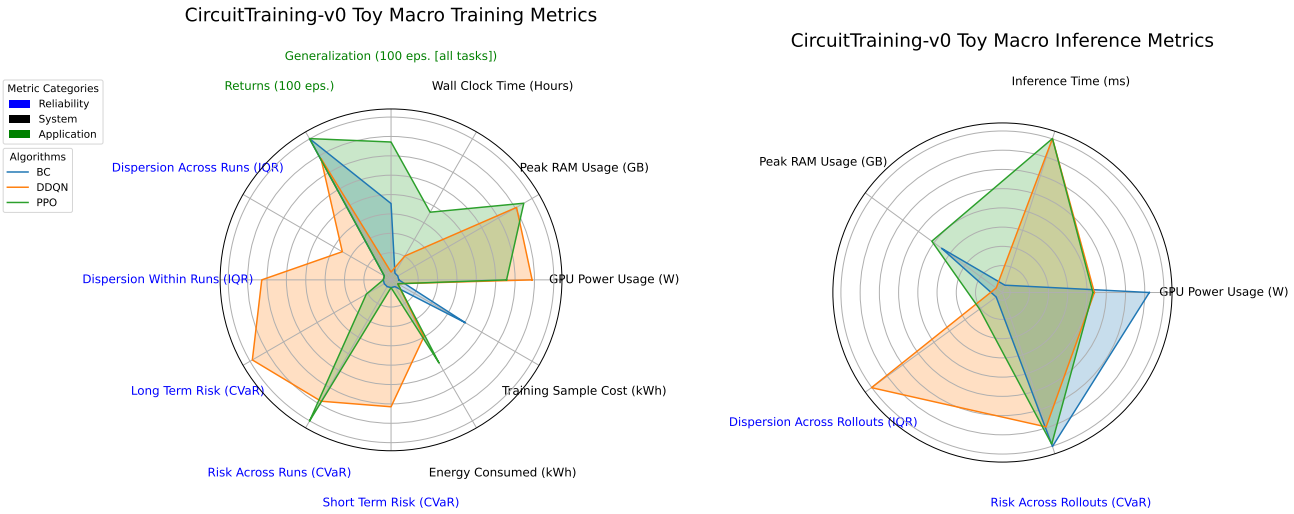

Figure 8: Graphical representation of metrics for the "Toy Macro" netlist task of CircuitTraining-v0

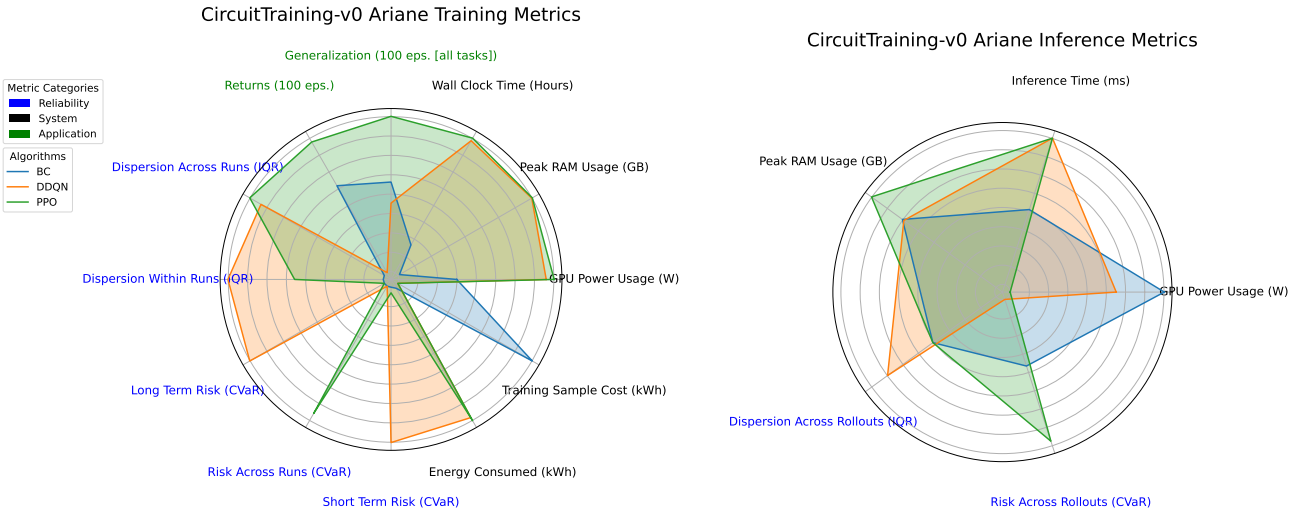

Figure 9: Graphical representation of metrics for the "Ariane" netlist task of CircuitTraining-v0

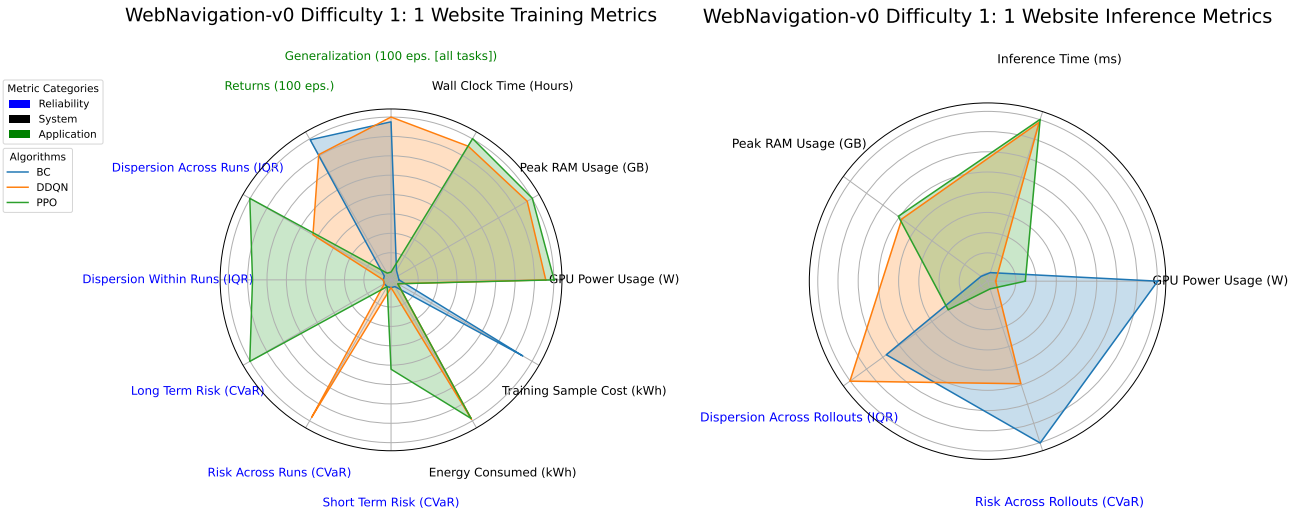

Figure 10: Graphical representation of metrics for the "difficulty 1, 1 website" task of WebNavigation-v0

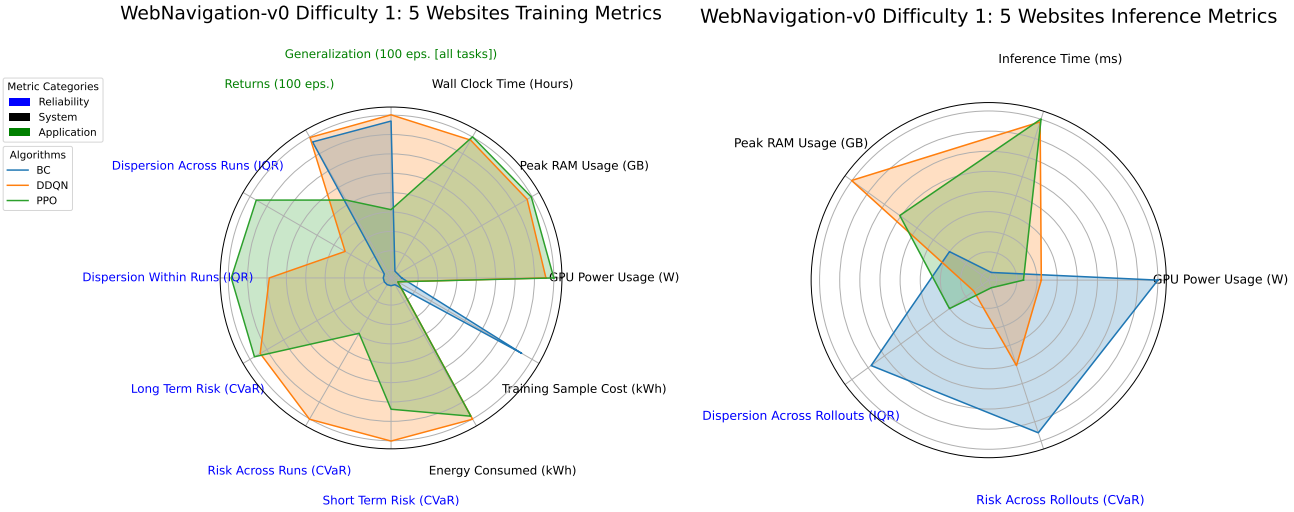

Figure 11: Graphical representation of metrics for the "difficulty 1, 5 websites" task of WebNavigation-v0

F. Experimental Setup

F.1. Training

We used the Tensorflow Agents (Guadarrama et al., 2018) library to conduct distributed reinforcement learning experiments across the three domains: computer chip floorplanning, web navigation, and quadruped locomotion. Our training setup consisted of one training server (a Google Cloud a2-highgpu-8g instance[1]) equipped with four NVIDIA A100 GPUs, and multiple collect servers (Google Cloud n2-standard-96 instances[2]) with 96 vCPUs running in parallel.

The number of collect jobs running simultaneously varied depending on the specific domain and the available resources (such as CPU and memory) on the collect machines, which are important for running the environments efficiently. When using a collect machine with 96 vCPUs, we adjusted the number of environment instances based on the computational requirements of each domain:

[1] cloud.google.com/compute/docs/gpus
[2] cloud.google.com/compute/docs/general-purpose-machines

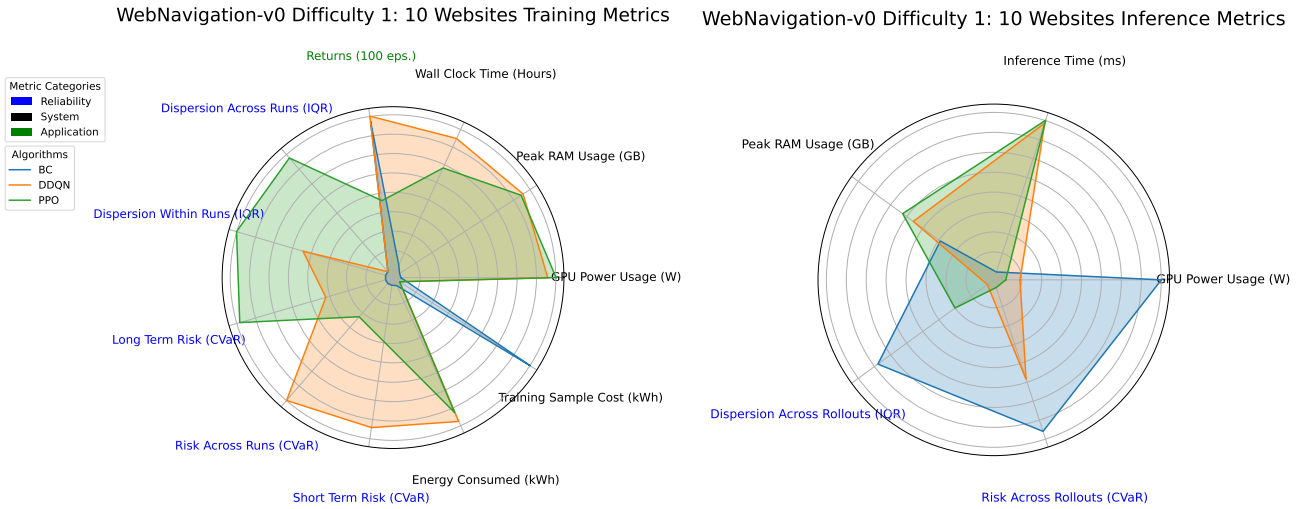

Figure 12: Graphical representation of metrics for the "difficulty 1, 10 websites" task of WebNavigation-v0

1. **Quadruped Locomotion**: With 96 vCPUs on the collect machine, we ran 44 quadruped locomotion environment instances concurrently using Python 3.9.
2. **Computer Chip Floorplanning**: For the computer chip floorplanning domain, we ran 25 computer chip floorplanning environment instances on a collect machine with 96 vCPUs using Python 3.10.
3. **Web Navigation**: When running web navigation experiments on a collect machine with 96 vCPUs, we instantiated 40 web navigation environment instances simultaneously using Python 3.10.

The behavioral cloning experiments for all three domains used the same setup as the online training experiments, with one training server equipped with four A100 GPUs.

F.2. Inference

For the inference phase, we used a single machine equipped with one NVIDIA V100 GPU to evaluate the trained models across all three domains: computer chip floorplanning, web navigation, and quadruped locomotion. The difference in hardware between the training and inference setups does not affect the application performance metrics, as these metrics are independent of the hardware and reflect the effectiveness of the trained models. However, the system performance metrics, such as inference time and memory usage, may vary depending on the specific hardware used during inference.

G. Hyperparameters

H. Dataset Collection

To collect datasets for each domain and task, we periodically saved the policies at fixed intervals throughout the training process. We then evaluated all the saved policies on 100 episodes for each domain and task. Based on these evaluations, we created a distribution of median returns and assigned an `expertise` level to each policy as follows:

1. `Novice`: The median return lies within one standard deviation below the mean.
2. `Intermediate`: The median return is within one standard deviation above or below the mean.
3. `Expert`: The median return is one standard deviation above the mean or higher.

In some cases, certain domains or tasks were too challenging, resulting in no policies of a given skill level. In such instances, we only provide a `novice` dataset.

I. Dataset Information

1. Dataset documentation and intended uses:
   - The A2Perf datasets consist of data collected from three simulated environments: computer chip floorplanning, web navigation, and quadruped locomotion. The data was generated by running reinforcement learning policies at various stages of training, capturing the experiences of these policies interacting with the respective environments. The datasets are intended for use in offline reinforcement learning, imitation learning, and hybrid approaches, allowing researchers to evaluate and compare different algorithms without the need for online data collection.
2. Dataset availability:
   - The datasets can be accessed at:
     - Circuit Training: `https://drive.google.com/drive/folders/1UMhLlnYmfbnjBPN_JwVy4YXDUahXrWf6`
     - Quadruped Locomotion: `https://drive.google.com/drive/folders/1n1BJFip-reSPif8Bv3jXAnSOgfQAEje7`
     - Web Navigation: `https://drive.google.com/drive/folders/13EmCscVatl7Q5EFdWFRpwKlA2yRfonE5`
3. Data format and usage:
   - The datasets are provided in the widely-used HDF5 format, a data model and file format designed for efficient storage and retrieval of large datasets. Detailed instructions on how to read and use the data with the Minari framework are provided at: `https://minari.farama.org/`
4. Licensing:
   - The A2Perf datasets are released under the MIT License. The authors confirm that they bear all responsibility in case of violation of rights.
5. Maintenance and long-term preservation:
   - The datasets are hosted on a Google Cloud Bucket maintained by the Farama Foundation, a non-profit organization dedicated to supporting open-source machine learning projects. This ensures the long-term availability and accessibility of the datasets for the research community.

J. Software Usage

A2Perf is a benchmark harness designed to be used flexibly on various machines. The user has the option to either run it in a Docker container or to run the benchmark locally. A Docker container is available in the harness and can be adapted to your needs. If you would like to run the benchmark locally, a guide is available to install the A2Perf benchmark harness on your Linux or MacOS system. While you can benchmark on both operating systems, it is important to note that system performance metrics are tracked using CodeCarbon. This allows to capture energy, power and memory usage at regular time intervals, and uses pyRAPL to compute the Running Average Power Limit (RAPL). However, RAPL uniquely measures power consumption information for Intel CPUs, DRAM for server architectures and GPU for client architectures. When using systems using CPU architectures different then the Intel CPUs, the power consumption metric will return a computed estimate rather than a measured metric.

The benchmark harness allows you to benchmark both the training and inference of your algorithm and agents respectively. In order to benchmark your algorithms, you need to create a submission folder which includes several files which A2Perf calls. First, a training file, `train.py` contains a function `train()`, which starts the training process of your algorithm when called. Similarly, `inference.py` covers the inference of your trained model. This file includes several functions responsible for the loading of your trained model, preprocessing observations and running inference on your model. Using a `requirements.txt` file, additional Python packages and versioning can be specified. Running the benchmark is done through a command line interface. Using flags, we can pass additional information to the submission to set up the benchmark. A `gin_config` flag allows the user to define the settings for your environment and training process. Additionally, we need to pass the path to the submission folder using the `participant_module_path` flag. For a more detailed description, tutorials are available in the repository.

K. Website Generation & Agent Interaction

To create environments for the web navigation tasks, we generate synthetic websites that agents must learn to navigate. These websites serve as training and evaluation environments, where agents need to fill forms and interact with various web elements. Here we describe our procedural website generation process.

To generate the set of websites W, we first assume a target number of websites, denoted as N_{websites}. Following the approach in Gur et al. (2021) (shown in Table 4 of the paper), we consider 42 possible primitives that can be added to a web page and introduce two additional primitives: a "new page" primitive and a "stop" primitive, resulting in a total of 44 primitives.

The website generation process begins with an empty web page. We repeatedly sample uniformly from the 44 primitives and add them to the current page. If the "new page" primitive is selected during the sampling process, we start adding primitives to a new linked page. If the "stop" primitive is selected, we conclude the generation of the current website and proceed to generate the next website, if necessary. This process continues until we have generated the desired number of websites, N_{websites}. Each website in the resulting set W consists of one or more web pages, with each page containing a sampled set of primitives.

We define the difficulty of a web page as the probability of a random agent interacting with the correct primitive(s). The difficulty of page p_i is given by $-\log\left(\frac{n_{\text{active}}}{n_{\text{active}}+n_{\text{passive}}}\right)$, where n_{active} and n_{passive} denote the number of active and passive primitives on the page, respectively. The difficulty of an entire sequence of web pages is determined by summing the difficulty of all individual pages it contains. Based on these difficulty calculations, we partition the websites into three difficulty levels. The three levels of difficulty correspond to the probability thresholds of 50%, 25%, and 10% for levels 1, 2, and 3, respectively. Users can select a specific difficulty level of web navigation by executing Python commands such as `env = gym.make("WebNavigation-Difficulty-01-v0", num_websites=1)`, where the `num_websites` argument defines the pool of websites available for the environment. During training or evaluation, each episode begins by randomly selecting one website from this pool at the specified difficulty level. For example, if `num_websites=10`, the environment will generate 10 websites at the specified difficulty level, and each episode will randomly assign one of these websites for the agent to navigate. At each timestep, the agent can interact with an HTML element on the page, such as modifying the text field or clicking on the element, with the objective of entering correct information into forms and clicking "next" or "submit" to advance between web pages.

L. Reliability Metrics

[14] https://en.wikipedia.org/wiki/Expected_shortfall

Phase	Metric Name	Description	Equation
Training	Dispersion Within Runs	Measures higher-frequency variability using IQR within a sliding window along the detrended training curve. Lower values indicate more stable performance.	$\frac{1}{T-4}\sum_{t=3}^{T-2}\text{IQR}\left(\{\Delta P_{t'}\}_{t'=t-2}^{t+2}\right)$
	Short-term Risk (CVaR)	Estimates extreme short-term performance drops. Lower values indicate less risk of sudden drops.	$\text{CVaR}_\alpha\left(\Delta P_t\right)_{t=1}^{T}$
	Long-term Risk (CVaR)	Captures potential for long-term performance decrease. Lower values indicate less risk of degradation.	$\text{CVaR}_\alpha\left(\max_{t'\leq t} P_{t'} - P_t\right)$
	Dispersion Across Runs	Measures variance across training runs. Lower values indicate more consistent performance across runs.	$\frac{1}{T}\sum_{t=1}^{T}\text{IQR}\left(\{P_{t,j}\}_{j=1}^{n}\right)$
	Risk Across Runs (CVaR)	Measures expected performance of worst-performing agents. Higher values indicate better worst-case performance.	$\text{CVaR}_\alpha\left(P_{T,j}\right)_{j=1}^{n}$
Inference	Dispersion Across Rollouts	Measures variability in performance across multiple rollouts. Lower values indicate more consistent performance.	$\text{IQR}\left(R_i\right)_{i=1}^{m}$
	Risk Across Rollouts (CVaR)	Measures worst-case performance during inference. Higher values indicate better worst-case performance.	$\text{CVaR}_\alpha\left(R_i\right)_{i=1}^{m}$

Table 16: Reliability Metrics from Chan et al. (2019) with Mathematical Formulations. P_t: performance at time t. $P_{t,j}$: performance at time t for run j. R_i: performance during rollout i. $\Delta P_t = P_t - P_{t-1}$: performance change between consecutive time steps (detrended value). CVaR_α[14]: Conditional Value at Risk at level α. IQR: Inter-Quartile Range. Sliding window length is 5 time steps centered on t, calculated over all t from 3 to $T-2$ to ensure the window is valid. T: total number of time steps. n: number of runs (10 for our experiments). m: number of rollouts (100 for our experiments).