InstaDeep®

Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies

Deep Learning: Classics and Trends (ML Collective) - 27 February 2026

Felix Chalumeau*, Daniel Rajaonarivonivelomanantsoa*, Ruan de Kock*,
Claude Formanek, Sasha Abramowitz, Omayma Mahjoub, Wiem Khlifi, Simon Du Toit,
Louay Ben Nessir, Refiloe Shabe, Arnol Fokam, Siddarth Singh, Ulrich Mbou Sob, Arnu Pretorius

InstaDeep RL Research team

Reinforcement Learning: Challenges of Zero-Shot Performance

- **RL** has achieved **groundbreaking** results: e.g. Atari, AlphaGo, ChatGPT.
- Despite years of improvements, many tasks are still **far from being solved** using RL.
- Sources of **complexity** (amongst others):
  - Large observation and action spaces (combinatorial)
  - Partial observability
  - Distribution shift between training and inference
  - Hard exploration, non-stationarity, etc...
- Numerous tasks are **not solvable on the first attempt** (zero-shot) and the gap to optimality can be very large.

RL's zero-shot performance ceiling.

Image Courtesy: Generated by ChatGPT4o

Demonstrating the Performance Ceiling

- We looked at a **reference benchmark** for RL (based on Dec-POMDPs)
- We observed that current SOTA was still **very far from optimal** on 35% of the 45 tasks
- To validate the existence of a performance ceiling, we **re-trained** that SOTA policy **until convergence**
- It still **plateaus below 70%** of the estimated max performance
- And a **third** of the scenarios remain under **50%**

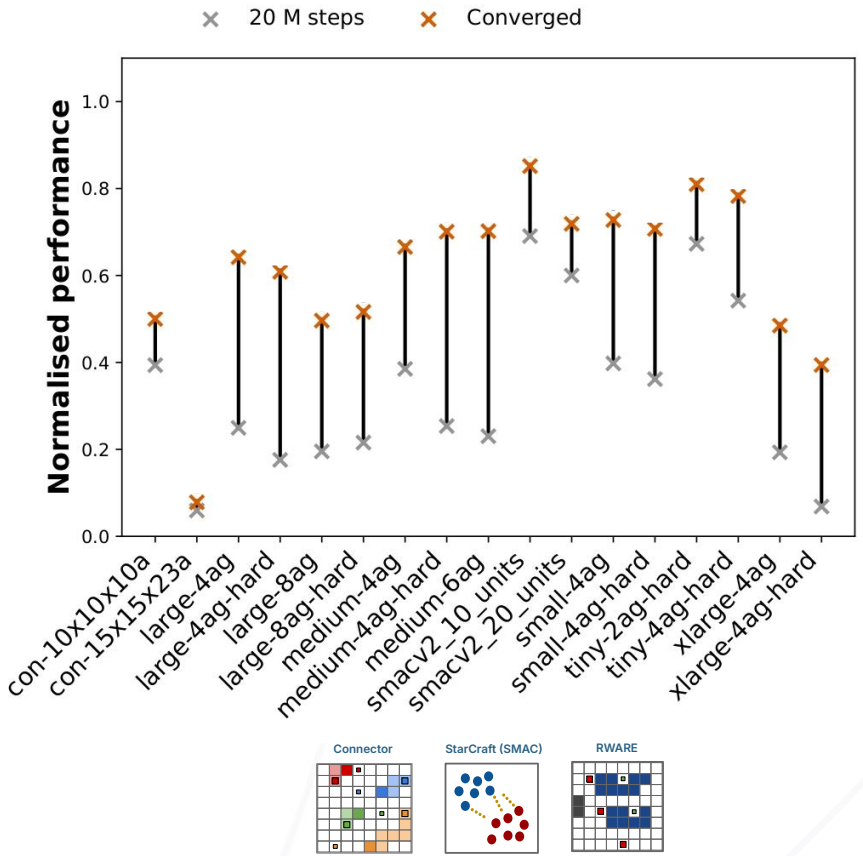


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Demonstrating the Performance Ceiling

- We looked at a **reference benchmark** for RL (based on Dec-POMDPs)
- We observed that current SOTA was still **very far from optimal** on 35% of the 45 tasks
- To validate the existence of a performance ceiling, we **re-trained** that SOTA policy **until convergence**
- It still **plateaus below 70%** of the estimated max performance
- And a **third** of the scenarios remain under **50%**

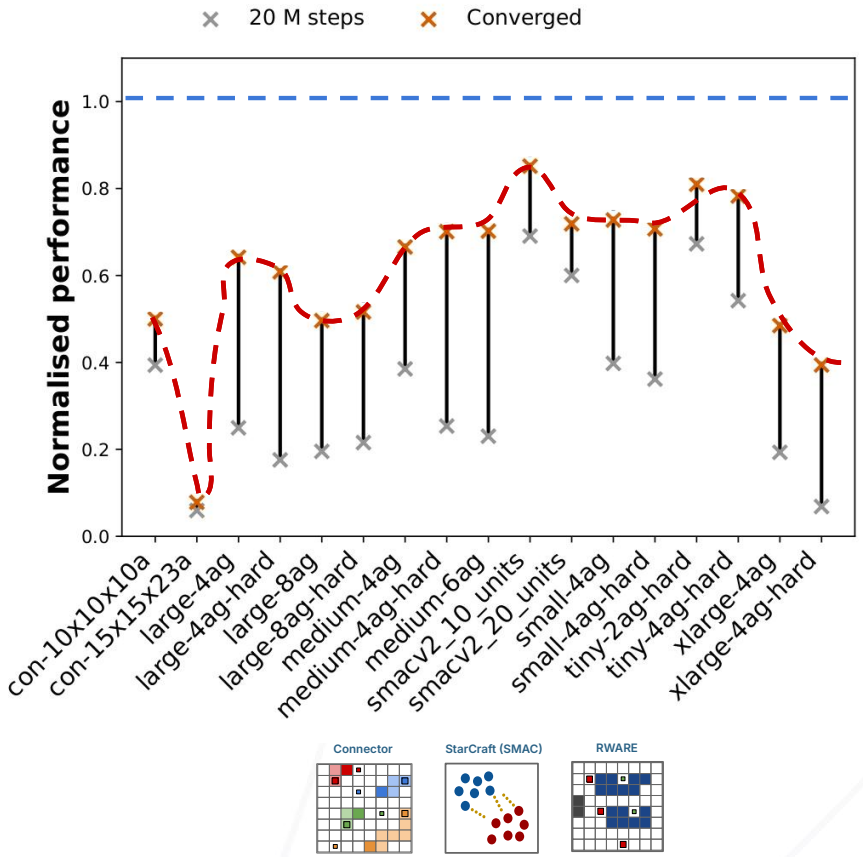


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

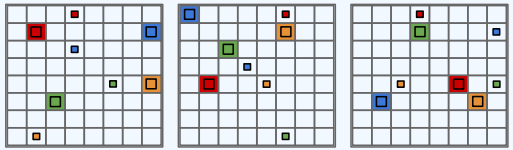
InstaDeep®

Problem Setting

The Reinforcement Learning Pipeline

1. TRAINING PHASE

Training distribution



Decision-making policy

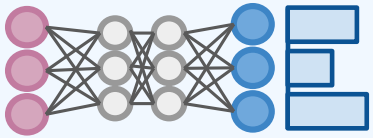
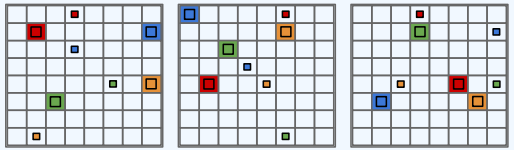


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

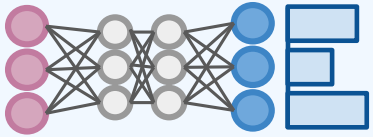
The Reinforcement Learning Pipeline

1. TRAINING PHASE

Training distribution

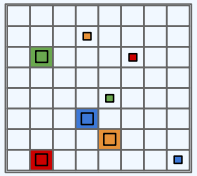


Decision-making policy

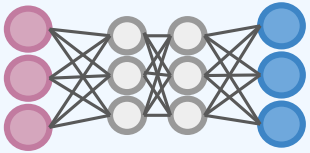


2. INFERENCE PHASE

New instance



Trained policy



Output a solution to the problem instance

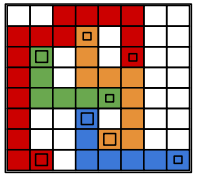
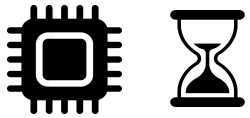


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Inference-Time Search: A Lost Opportunity

- Often, a **time** and **compute** budget is **available**, it is hence possible to get **several attempts** to the same problem before submitting a final **solution**
- Furthermore, we have access to a **scoring** function (or even a reward model) to evaluate our attempts
- Hence, one can use an **"inference-time strategy"** to help finding the **best possible solution** to the problem given a **time constraint** and a **compute capacity**

Compute & Time budget



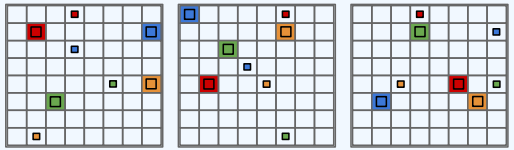
Can (must!) be used for search

Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

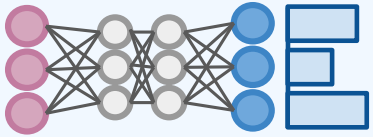
Inference-Time Search: Problem Setting

1. TRAINING PHASE

Training distribution

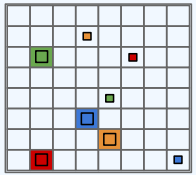


Decision-making policy

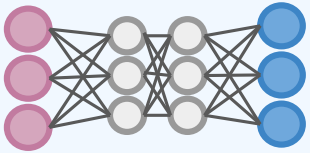


2. INFERENCE PHASE

New instance



Trained policy



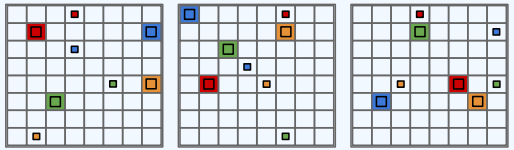
Compute & Time budget

Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

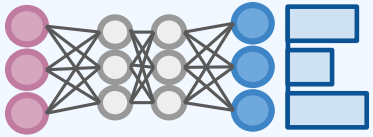
Inference-Time Search: Problem Setting

1. TRAINING PHASE

Training distribution

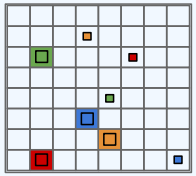


Decision-making policy

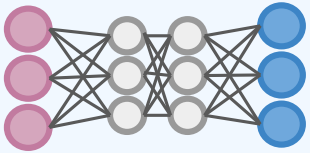


2. INFERENCE PHASE

New instance



Trained policy



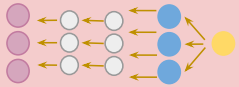
Compute & Time budget

Inference strategies

Stochastic sampling

Tree search

Online fine-tuning



COMPASS' search

Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

InstaDeep®

Experiments

Experimental Setting

- We use tasks with **multiple sources of complexity** (high dimensionality, partial observability, ...)
- We select the **17 most complex tasks** from a reference Multi-Agent RL benchmark

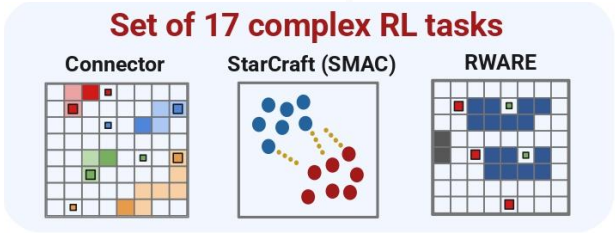


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Experimental Setting

- We use tasks with **multiple sources of complexity** (high dimensionality, partial observability, ...)
- We select the **17 most complex tasks** from a reference Multi-Agent RL benchmark
- We use **3 base policy architectures:** 2 most common in MARL (IPPO, MAPPO), and SOTA zero-shot (Sable).
- Using a broad set of **time and compute** budget

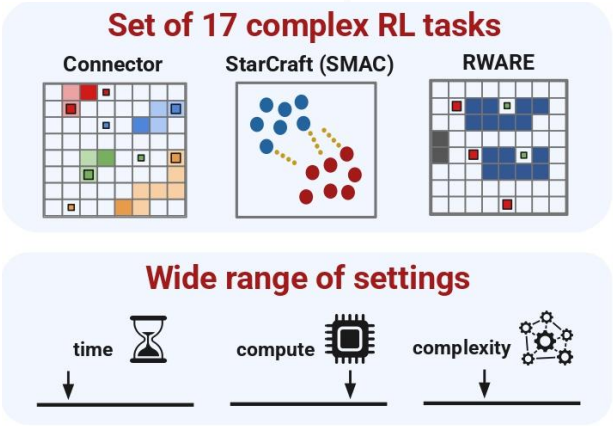


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Experimental Setting

- We use tasks with **multiple sources of complexity** (high dimensionality, partial observability, ...)
- We select the **17 most complex tasks** from a reference Multi-Agent RL benchmark
- We use **3 base policy architectures:** 2 most common in MARL (IPPO, MAPPO), and SOTA zero-shot (Sable).
- Using a broad set of **time and compute** budget
- Evaluate **4 types** of **inference strategies**

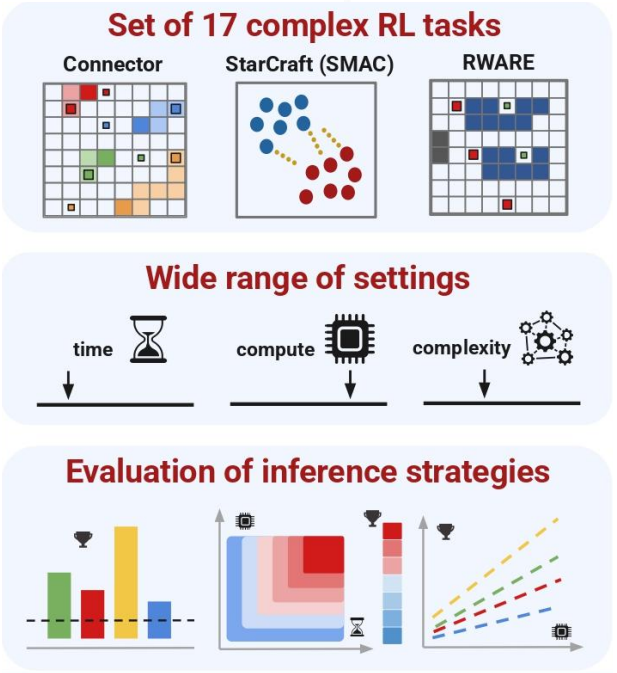


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

InstaDeep®

Can We Have a High Impact on Performance within a Few Seconds of Search?

Few-Seconds Performance: Experimental Setting

Main design choices:

- We represent the **compute capacity** by the number of parallel attempts one can get, and fix it to 64
- We allow a **time budget** of **30 seconds**
- We evaluate the **3** base **policies**, with each **4** inference **strategies** on all **17 tasks** (128 seeds)
- We report the inter-quartile mean over tasks with 95% stratified bootstrap confidence intervals

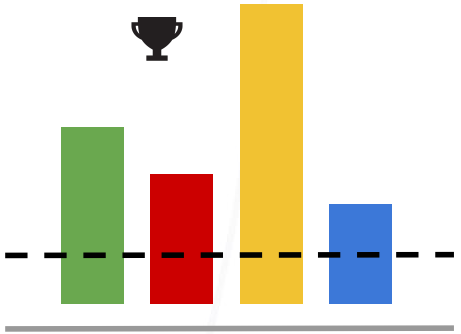


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Few-Seconds Performance: Results

Main observations:

- **Each base policies** gets a **significant boost** from using an inference-time **search**
- Good search + Bad policy > Zero-shot SOTA
- **COMPASS** is the **leading** inference strategy across policies and tasks
- The performance **gain** of the leading inference-time method **over zero-shot** is massive (~ **+45%**)

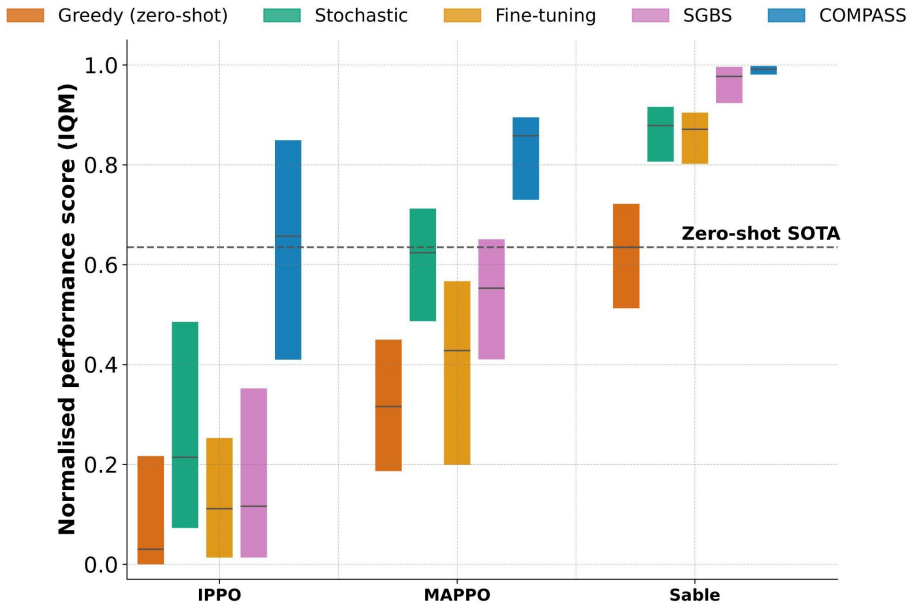


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Few-Seconds COMPASS to Break the Ceiling

Using inference-time search enables to push all results **above 80%** within a couple seconds

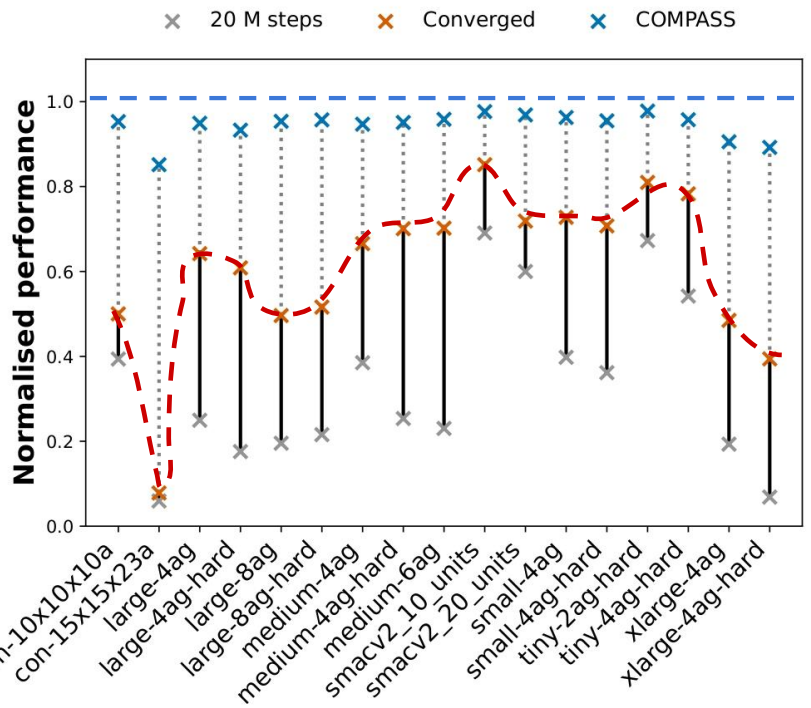


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Few-Seconds COMPASS to Break the Ceiling

... reaching **up to 126%** improvement on the hardest task.

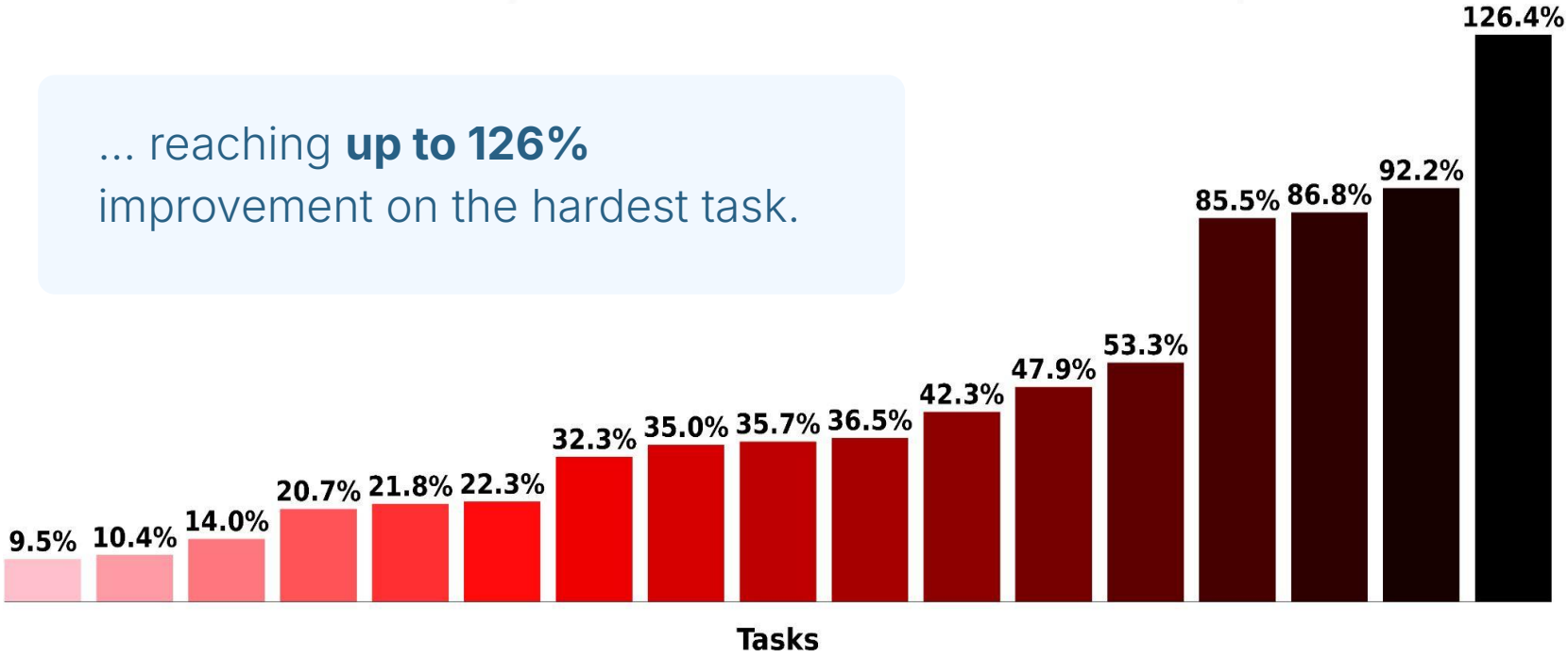


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

InstaDeep®

How Do Methods Compare in Various Time and Compute Settings?

Scaling Time and Compute: Experimental Setting

Main design choices:

- Most papers in the literature focus on a very **precise setting**
- **Comparative performance** of inference strategies is **impacted by the budget** available
- We evaluate the best base policy (Sable) for **300 seconds** on **compute** capacities going from **4 to 256**
- We report **performance contour plots** obtained for each time and compute capacity

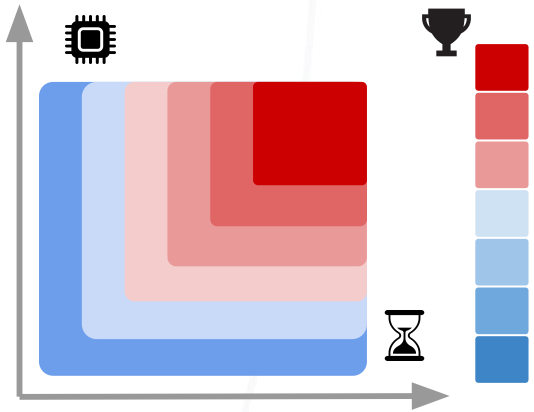


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Scaling Time and Compute: Results

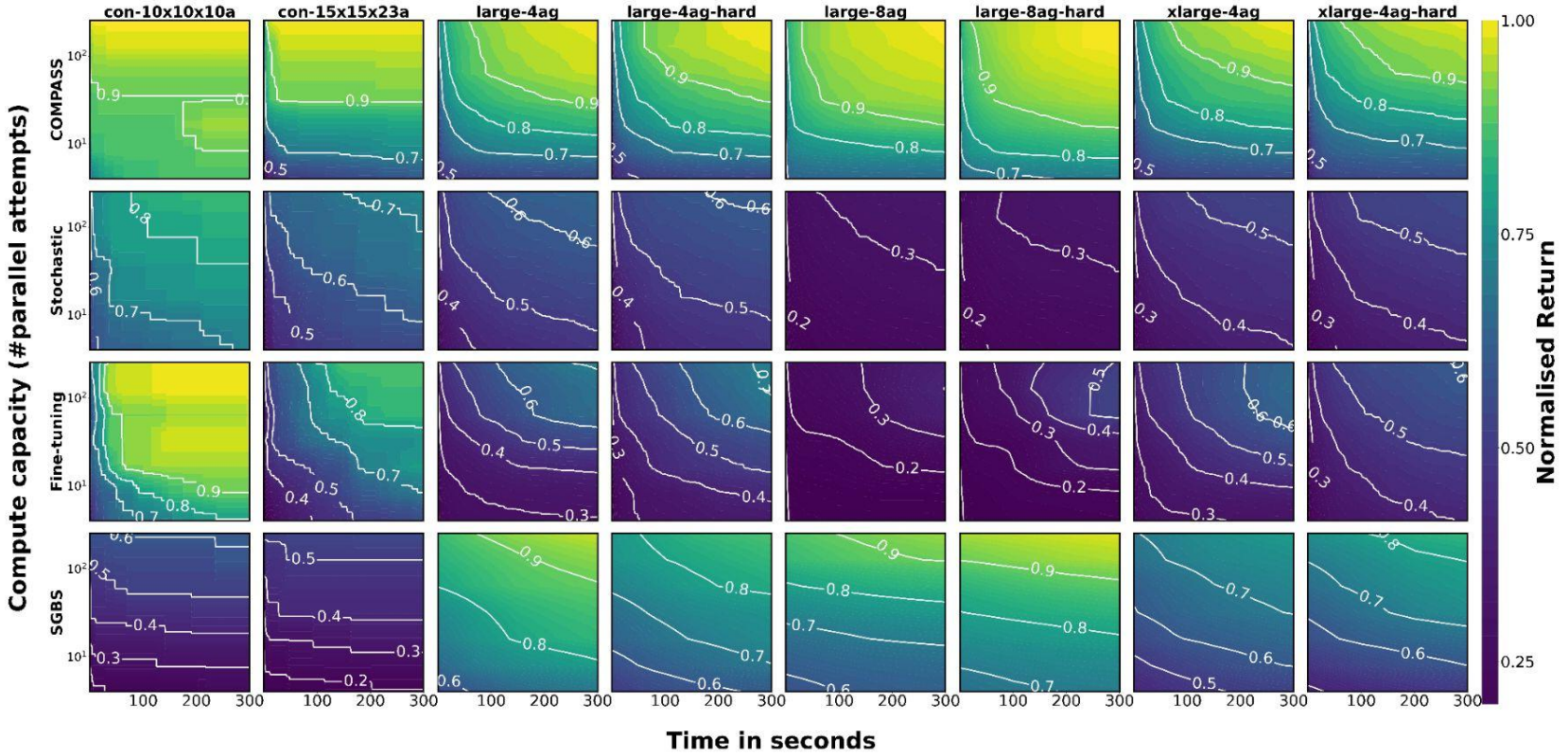


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Scaling Time and Compute: Results

Main observations:

- All methods **benefit from increased budgets**
- **Online fine-tuning** suffers from high variance
- **The diversity-based approaches (COMPASS)** is consistently the best
- **Tree search** can be better than COMPASS in some low-compute settings

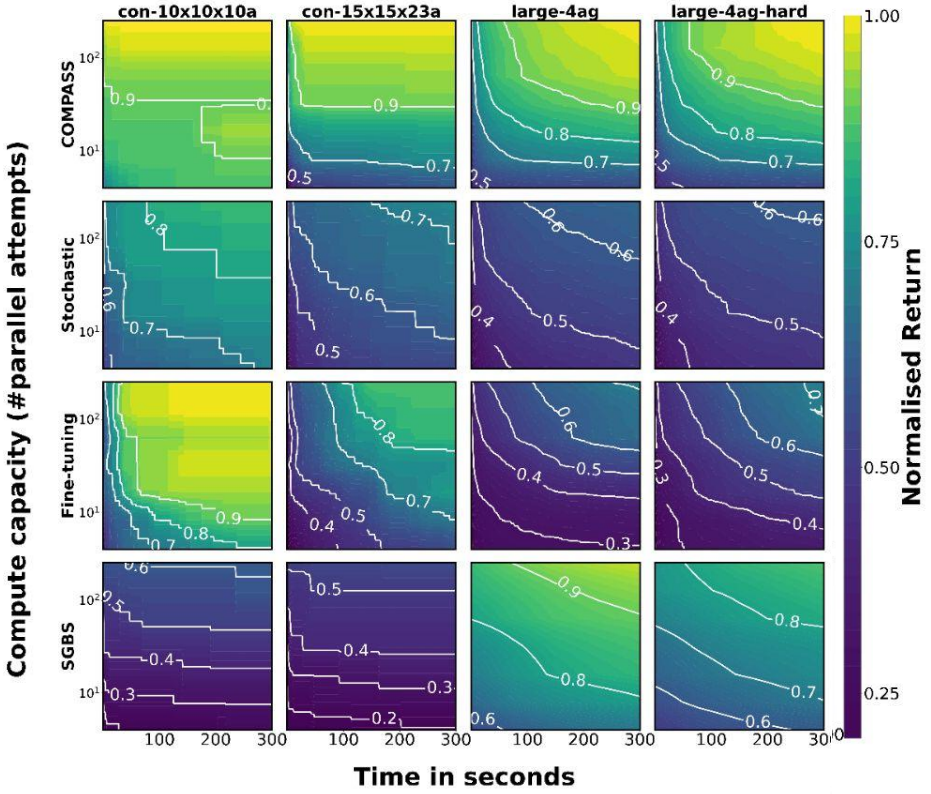


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

InstaDeep®

How Well do Methods Scale with Compute?

An Alternative View: Compute Scaling Trends

Main design choices:

- In practice: **time** is often **limited**
- But **compute** can be **scaled**
- We keep time fixed, and we **increase the compute capacity** available to the methods

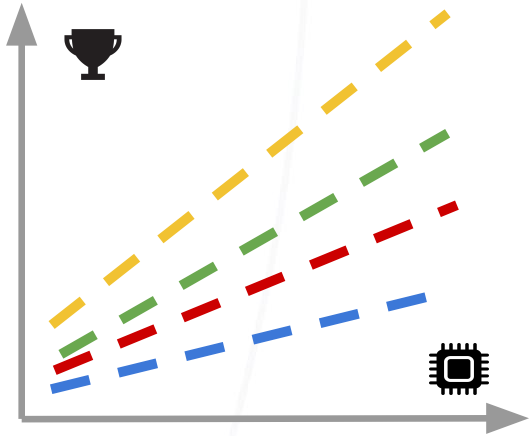


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Compute Scaling Trends: Results

Main observations:

- Stochastic sampling scales the least (expected)
- Fine-tuning needs compute capacity to stabilise gradient but is not competitive on this benchmark **(refuting the misbelief that overfitting to an instance is enough)**
- The tree search (SGBS) is more **robust** than online fine-tuning, providing a great **off-the-shelf strategy**
- COMPASS scales very well with compute (thanks to diversity and robust adaptation)

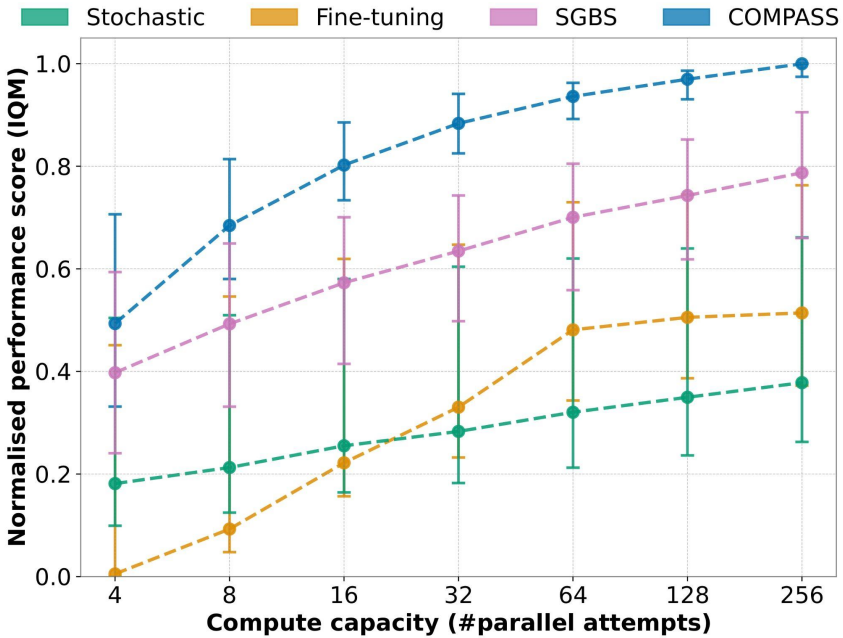


Image Courtesy: Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies (Chalumeau et al., 2025)

Take Home Messages

- In RL tasks, **zero-shot** performance can **reach a ceiling**, far away from the optimal achievable solution
- In many real-world situations, using **inference-time search is possible**, but practitioners tend to ignore it
- Inference-time search is **key to breaking this performance ceiling** and producing competitive solutions to complex problems

InstaDeep®

Thank You