No reward? No problem: Training Maze Agents with Preferences

Giovanni Filomeno
Supervisor: Vihang Patil

Motivation:

Traditional reinforcement learning relies on dense, well-designed rewards, but these are hard to define in complex environments like mazes.

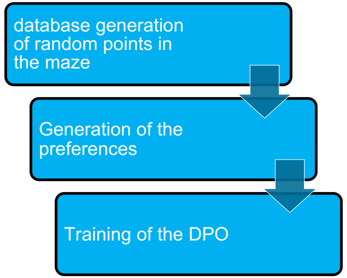
Challenge:

How can we train an agent to navigate effectively without access to a handcrafted reward function?

Idea:

Use **Direct Preference Optimization (DPO)** to teach the agent via comparisons between good and bad trajectories, learning a policy that reflects desired behaviors.

Workflow:

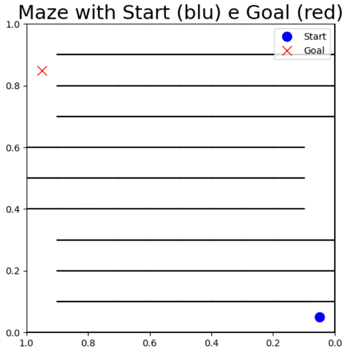


Breaking:

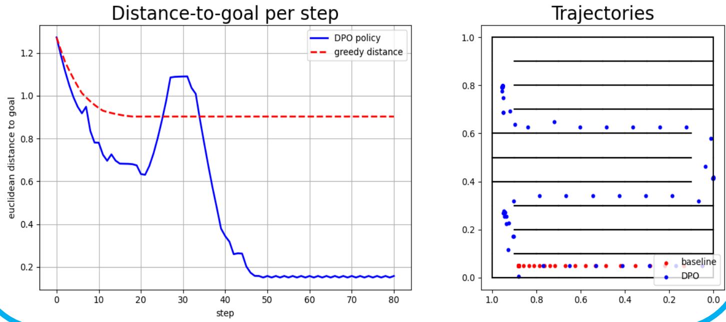
- **Environment**: I designed a custom 2D maze environment with winding "S-shaped" corridors that challenge naive strategies.
- **Point Generation**: I generated thousands of random points (states) within the Maze
- **Preference**: For each pair of points, I computed a custom preference score based on: distance to goal, distance to wall, death end, real distance
- **Learning with DPO:** A neural network was trained to model preferences, learning to assign higher scores to better states. The training objective was to rank preferred points higher — effectively shaping a reward-free value function.
- **Sensitivity (Future Work):** The final policy was heavily influenced by the way we designed the scoring function. Small changes in how we penalized dead-ends or long paths led to drastically different behaviors, highlighting the importance of well-designed preferences.

Introduction:

- In many real-world tasks, defining a precise reward function is difficult or even misleading.
- Preference-based learning offers a flexible alternative: agents learn by comparing which behaviors are better.
- Direct Preference Optimization (DPO) is a scalable method to learn from pairwise preferences without explicit rewards.
- We apply DPO to train agents in 2D maze environments, learning to navigate effectively using only trajectory comparisons.



Results: DPO can escape from Maze



JOHANNES KEPLER UNIVERSITY LINZ