Overview

WebShop là một môi trường trang web thương mại điện tử mô phỏng với 1,18 triệu sản phẩm thực tế và 12.087 hướng dẫn văn bản do cộng đồng đóng góp. Trong môi trường này, một tác nhân cần điều hướng nhiều loại trang web và đưa ra các hành động đa dạng để tìm kiếm, tùy chỉnh và mua sản phẩm dựa trên hướng dẫn

Sau khi có danh sách sản phẩm:

* Agent **đọc mô tả, so sánh các sản phẩm**, đánh giá độ phù hợp với instruction.
* Có thể **click vào sản phẩm để xem chi tiết**, lọc kết quả theo giá, đánh giá, v.v.

1. **Imitation Learning (Behavior Cloning - BC)**

* **Imitation Learning** là quá trình huấn luyện một tác tử (agent) để bắt chước **chính sách hành động** (policy) của **expert (chuyên gia)** thông qua **các ví dụ (data)** thay vì thông qua trải nghiệm phần thưởng. rewward

🔸 **Ý tưởng:**  
Huấn luyện agent **bắt chước hành vi con người** (từ dữ liệu người dùng đã thực hiện nhiệm vụ mua sắm). Agent học cách:

* Từ một instruction và danh sách sản phẩm,
* Chọn sản phẩm **giống như con người đã chọn** trong tình huống tương tự.

🔸 **Cách hoạt động:**

* Dữ liệu huấn luyện chứa (instruction, list of products, action)
* Sử dụng model như BERT hoặc LSTM để encode instruction + sản phẩm
* Huấn luyện theo cách supervised learning → học chính xác hành động con người đã làm

🔸 **Ưu điểm:**

* Học nhanh, không cần reward signal
* Dễ huấn luyện nếu có dữ liệu đủ

🔸 **Nhược điểm:**

* Không tự khám phá
* Phụ thuộc vào chất lượng hành vi trong dữ liệu

Công thức:

Hành động rời rạc(Cross-Entropy Loss) Khi agent chỉ được **chọn một trong số lượng hữu hạn các hành động rõ ràng**, ví dụ như:

 Đi **trái / phải / lên / xuống**

 **Mua / không mua**

 **Bấm nút A / B / C**

 **Chọn sản phẩm số 1, 2, 3,...**

Hành động liên tục Khi agent được chọn hành động là một hoặc nhiều giá trị thực (số thực) như:

Góc quay

Tốc độ

Lực tác động

Vị trí x, y

## ✅ 2. ****Reinforcement Learning (PPO – Proximal Policy Optimization)****

🔸 **Ý tưởng:**  
Huấn luyện agent thông qua **thử nghiệm – phản hồi – học từ reward**:

* Agent thử nhiều hành động (click, tìm kiếm, chọn sản phẩm)
* Nhận reward nếu chọn đúng sản phẩm phù hợp
* Tối ưu hóa chính sách hành động để nhận nhiều reward hơn

🔸 **Cách hoạt động:**

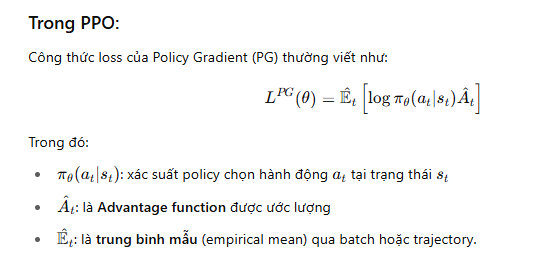
* Encode instruction + sản phẩm → đưa vào policy network
* PPO cập nhật chính sách để tối đa hóa reward dài hạn
* Agent học cách **phân biệt sản phẩm tốt / không phù hợp** dần theo thời gian

🔸 **Ưu điểm:**

* Có thể vượt qua chất lượng hành vi ban đầu
* Phù hợp khi cần khám phá

🔸 **Nhược điểm:**

* Cần nhiều lượt chơi (interaction) để học tốt
* Chậm hơn imitation learning



trung bình thực nghiệm ko phải trung bình mẫu

**E^t​** là **trung bình thực nghiệm tại thời điểm t**, dùng để xấp xỉ kỳ vọng trong loss function của PPO.

