1) Episode cumulative reward（訓練/評估：每回合累積 reward 曲線）

顯示內容：每一個 episode（或評估回合）結束時的累積 reward（縱軸），橫軸為 episode 編號。通常用來觀察智能體是否隨著訓練逐步改善。

代表意義：

- reward 是你在環境中定義的目標函數（本案為 -(α * emissions + β * commute_time)），數值越高代表「全體目標越好」（通勤時間和排放總和越小）。
- 若曲線呈現上升趨勢 → 智能體正在學習到更好的策略。
- 若曲線平坦或下降 → 可能學習停滯、回饋設計有問題或超參數需要調整。

如何判讀 / 偵錯：

- 初期高度波動是正常（探索造成），但長期應趨穩定。若長期波動很大，考慮減少探索率或增加 training timesteps。
- 若 reward 迅速飽和在很差的值，檢查 reward scale（α, β）是否不合理，或 reward 是否帶有常數偏移（例如通勤時間基準過大）。
- 若 reward 上升但通勤或排放某一項惡化，表示 reward 權重需要重新設計（檢視個別指標）。

---

2) Final commute time（每回合結束的通勤時間）

顯示內容：每一個 episode 結束時計算的通勤時間（或平均通勤時間），縱軸為通勤時間、橫軸為 episode。

代表意義：

- 直接衡量交通效能（通勤延遲、交通擁塞程度的 proxy）。
- 配合 reward 曲線，可看智能體是否用犧牲能源或其他指標來換取較短通勤時間（trade-off）。

如何判讀 / 偵錯：

- 若通勤時間下降但 emissions 上升，代表策略偏向犧牲環境換取效率（檢查是否需要引入公平或成本約束）。
- 若通勤時間無明顯改善，但 reward 提升，檢視是不是 emissions 減少造成 reward 上升（要解釋是如何 trade-off 的）。

---

3) Final emissions（每回合結束的排放量）

顯示內容：每回合結束時計算的 emissions（縱軸），橫軸為 episode。

代表意義：

- 衡量能源使用與污染（由 environment 的 energy_demand 與 traffic_load 組成）。
- 與通勤時間一起觀察可以直接看出「政策 trade-off」。

如何判讀 / 偵錯：

- emissions 與 commute_time 若同時下降 → 理想情況（策略雙贏）。
- 若 emissions 波動很大，可能是環境噪聲太強或 agent 策略對 stochasticity 敏感 → 建議多 seed 重複實驗檢查穩健性。
- 若 emissions 下降但 reward 未提升太多，可能因為 α 權重過小（檢查 reward 公式）。

成功／失敗的具體判準（指標化）

- **成功-like**：相較 baseline，agent 在多個 seed 下平均能降低累積 cost（reward 提升），且通勤與排放至少一項顯著改善且另一項不惡化（或有可接受 trade-off）。
- **問題警訊**：
  - reward 與某一目標方向衝突（例如 reward 提高但通勤時間惡化）→ 檢查 reward 權重與指標計算。
  - 高 variance across seeds → 模型不穩定或 environment 過噪（考慮增加訓練時間、穩定化演算法或希臘式探索策略）。
  - action collapse（agent 始終選同一動作）→ reward 信號可能過弱或探索不足。