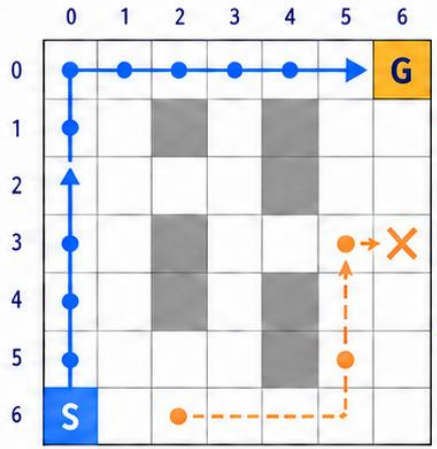
From environment contract to evidence

Environment contract

4 bounded TODOs · 13 checks

GridRoom (7 × 7)



Q-learning and policy

Q(s, a)

	↑	→	↓	←
S_0	…	…	…	…
S_1	…	…	…	…
S_2	…	…	…	…
:				
S_n	…	…	…	…

Baseline comparison

Random

Shortest path

Learned greedy

Export policy artifact

```
schema_version:
gridroom-policy.v1
q_values
[ ... ]
```

SHARED REHEARSAL · EVERY STUDENT

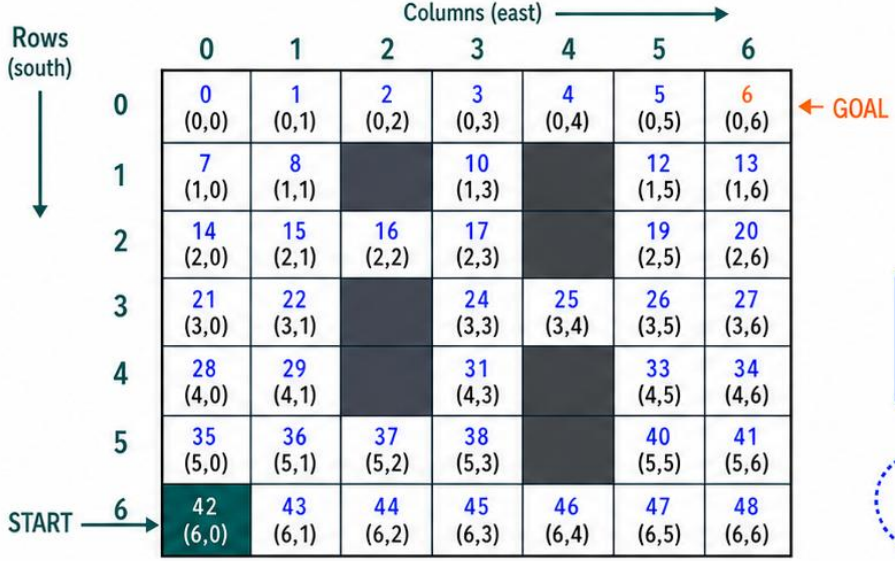
A transparent Q-learning lab for virtual worlds

Python + Gymnasium + NumPy + Matplotlib

4 bounded TODOs · 13 checks

The policy is not the result. The evidence package is.

The GridRoom contract is fixed



Blocked cells (7)

(1,2), (1,4), (2,4), (3,2), (4,2), (4,4), (5,4)

Observation

row-major state id

s = row × 7 + column

Collision / invalid move

self-transition

collision = true

start (6,0) · goal (0,6)

S observation = row × 7 + column

0 north · 1 east · 2 south · 3 west

R goal +10 · step −0.05 · collision −0.20 additional

40 steps: external truncation

reset on either / bootstrap-mask only termination

Episode boundary semantics

TERMINATED

reached goal

bootstrap = false

TRUNCATED

40 steps: external truncation

bootstrap = true

ONGOING

neither condition

bootstrap = true

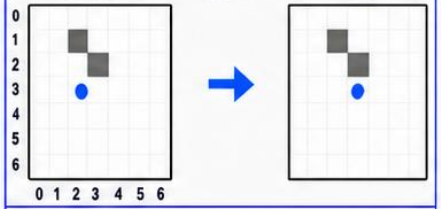
Deterministic transitions.

Invalid moves are self-transitions with collision = true.

Test dynamics before learning

valid move | wall / obstacle → self-transition + collision | goal → terminated | step limit → truncated | episode ended = terminated OR truncated | continuation zero = terminated

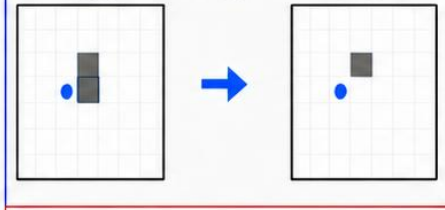
1 Valid move



Expected:
collision = False
terminated = False
truncated = False

PASS

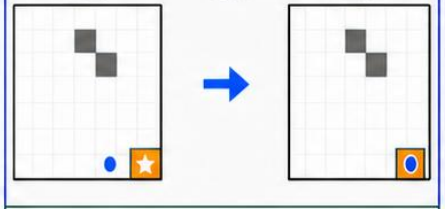
2 Wall / obstacle
self-transition + collision



Expected:
collision = True
terminated = False
truncated = False

PASS

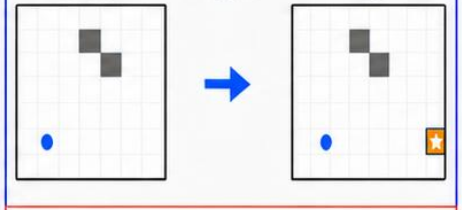
3 Goal
terminated



Expected:
collision = False
terminated = True
truncated = False

PASS

4 Step limit
truncated



Expected:
collision = False
terminated = False
truncated = True

PASS

Episode end (reset needed)

episode ended = terminated OR truncated

⟺

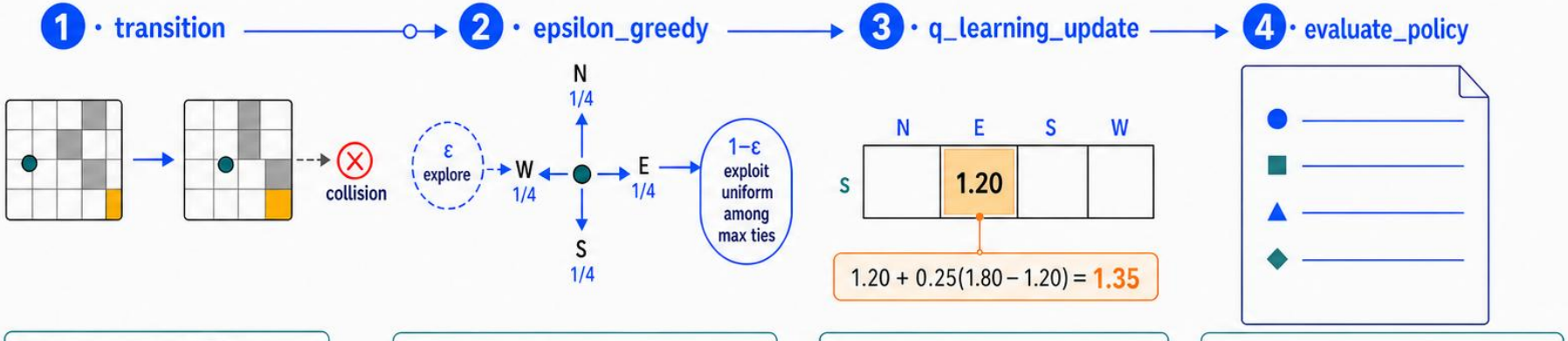
Q-Learning continuation (bootstrap mask)

continuation zero = terminated

■ = True
□ = False

A green curve cannot repair a wrong transition contract.

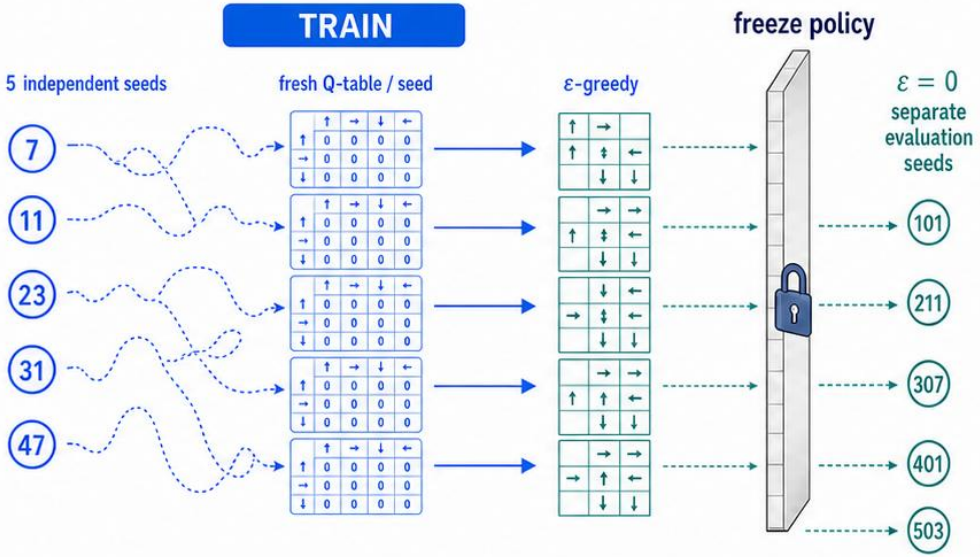
Four TODOs implement the learning spine

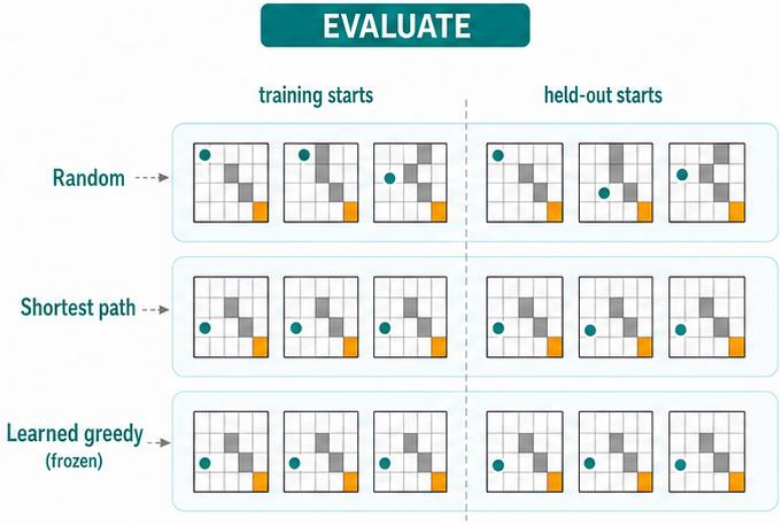


plotting | BFS baseline | aggregation | JSON export

Students edit one notebook.

Training is not evaluation





EVIDENCE (TRAINING STARTS)

Shortest path
100% success | 10.80 mean steps

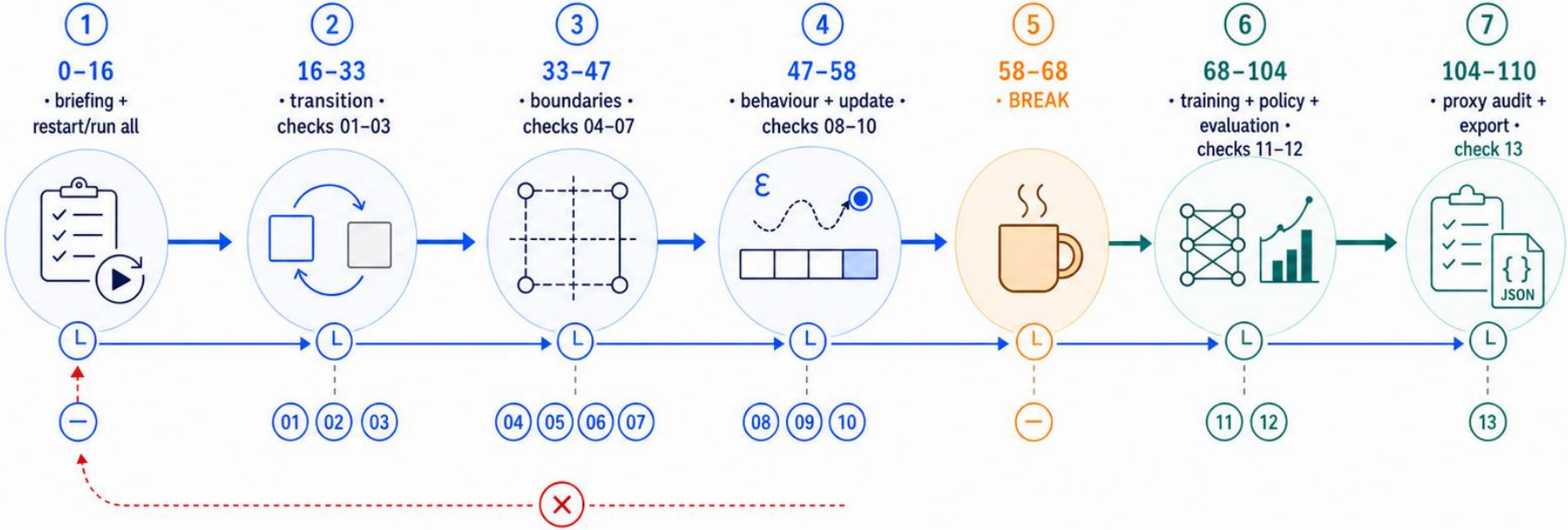
Learned greedy
100% success | 10.96 mean steps

Random
10% success

The heuristic is allowed to beat learning.

Held out as initial conditions, not necessarily unseen states.

The 110-minute route

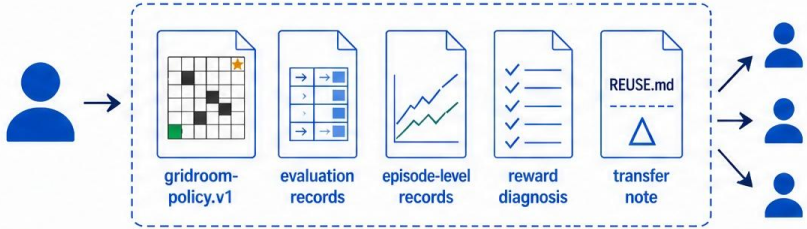


Fix the first broken contract, then restart and run all.

Finish the rehearsal; the project policy starts later

INDIVIDUAL EVIDENCE • 13/13 + personal policy/evaluation + transfer note

COMMON REFERENCE • identical verified instructor artifacts for everyone

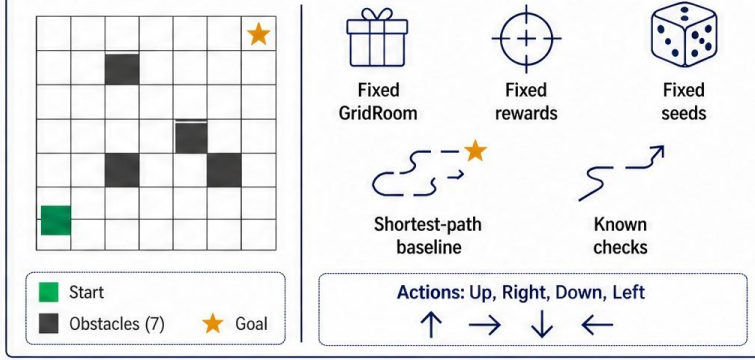


GUARDED BOUNDARY

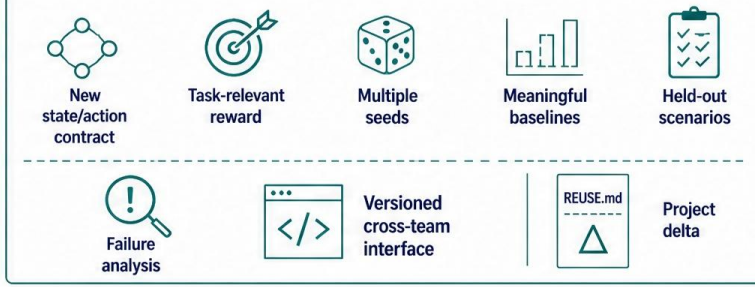
NO FINAL-PROJECT CREDIT

Complete together now; specialize later.

LAB • fixed GridRoom + rewards + seeds + baselines + known checks



PROJECT • new state/action + task reward + seeds + baselines + held-out + failures + interface



UNMODIFIED LAB SCAFFOLDING • NO FINAL-PROJECT CREDIT

Complete together now; specialize later.