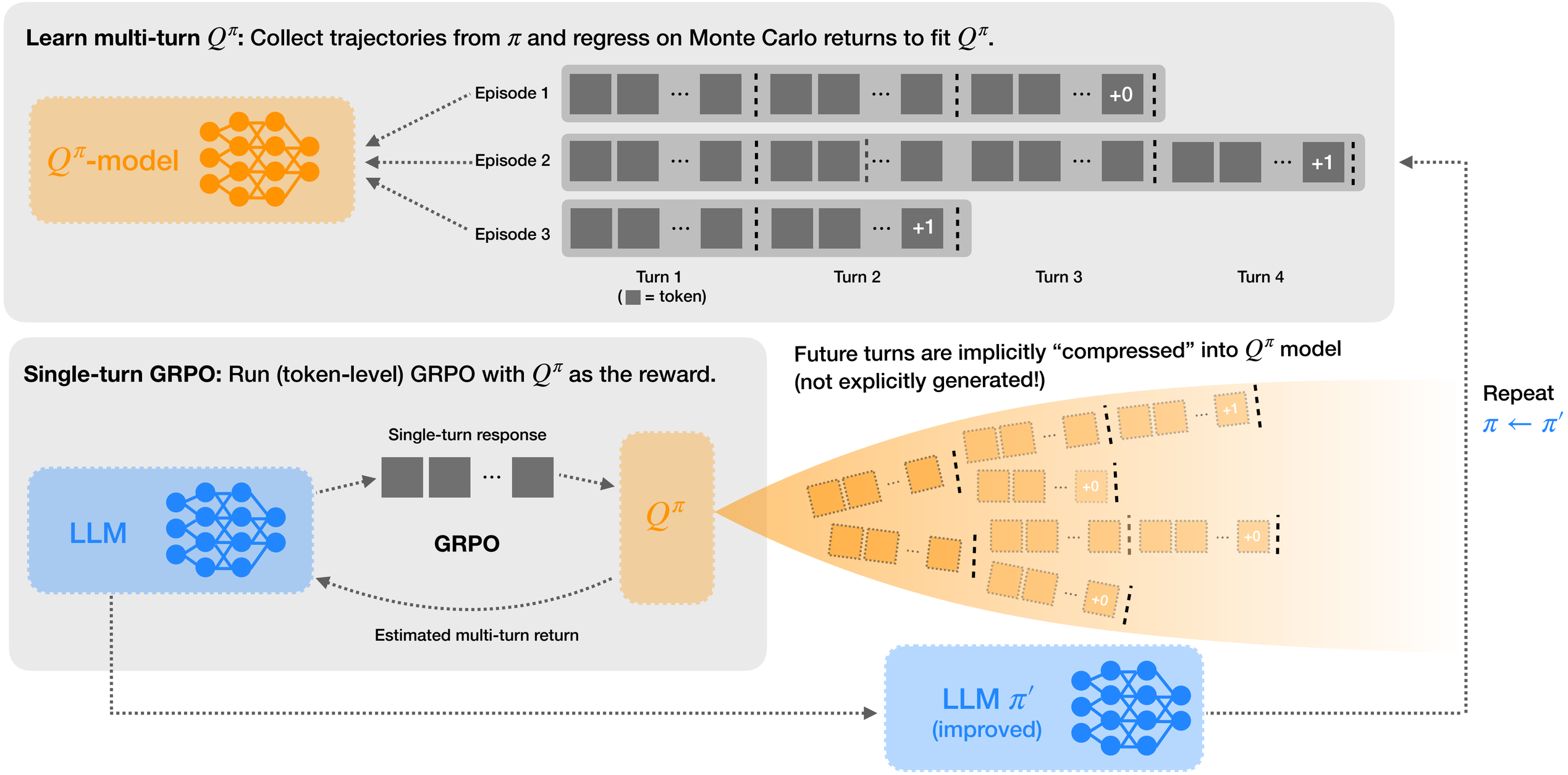

Learn multi-turn Q^π: Collect trajectories from π and regress on Monte Carlo returns to fit Q^π.
Q^π-model
Episode 1
Episode 2
Episode 3
+0
+1
+1
Turn 1
(= token)
Turn 2
Turn 3
Turn 4
Single-turn GRPO: Run (token-level) GRPO with Q^π as the reward.
LLM
Single-turn response
GRPO
Q^π
Estimated multi-turn return
Future turns are implicitly "compressed" into Q^π model
(not explicitly generated!)
+1
+0
+0
+0
Repeat
$\pi \leftarrow \pi'$
LLM π'
(improved)