Week 10 · Acts 2–3 — Refuse RL · Concept drift

Goal: (2) Compare one-shot SL vs online learning on static labels.
(3) See frozen SL fail on novel attack names.

Thesis: Adaptive ≠ reinforcement learning. Drift needs monitoring/retrain — not a Q-table on a CSV.

Run top to bottom. Keep the seen-vs-novel numbers for the exit ticket.

In [ ]:
```
import seas8414_teach as teach
teach.setup()
print(teach.__version__)
teach.story_title_card(
    "Acts 2–3 · Tool ladder + drift",
    "Wrong tool costs samples; frozen models miss novelty.",
    ["Act 2: one-shot SL vs online", "Act 3: seen vs novel recall"],
)
```

In [ ]:
```
refuse = teach.measure_rl_refusal()
print(refuse.sl_recall, refuse.online_final_recall, refuse.online_samples)
```

In [ ]:
```
teach.story_act2(refuse)
```

In [ ]:
```
drift = teach.measure_drift()
print(len(drift.unseen_attack_names), drift.seen_recall, drift.unseen_recall
```

In [ ]:
```
teach.story_act3(drift)
```

In [ ]:
```
# Optional: needs OPENAI_API_KEY in ~/.env — skip if unset
teach.ask(
    "Contrast Act 2 refusal vs Act 3 drift remedies in security English.",
    sl_recall=refuse.sl_recall,
    online=refuse.online_final_recall,
    seen=drift.seen_recall,
    novel=drift.unseen_recall,
)
```

Exit ticket

(a) Evidence RL is wrong for Act 1's task? (b) Best Act 3 number?