Corresponding essay: **E20** "You Cannot Benchmark a Coworker"

THE FOUR-QUESTION LAUNCH SCORECARD

Readiness Promotion Board

Convene a promotion board for one action class. Score it against the four questions, then decide whether to raise, hold or reduce its authority in the next release.

BRING One action class you want to make autonomous, and tasks drawn from your real policies

Score the four questions

Score each on 0-3: 0 not assessed, 1 evidence weak, 2 evidence partial, 3 evidence holds. Capability alone never clears the board.

Criterion	Score (0-3)	Evidence
Capability - the agent does the task on a clean, well-formed run with tools behaving	0 1 2 3	
Reliability - a stated pass^k target met on an evaluation set built from real and synthetic tasks, run many times each	0 1 2 3	
Autonomy fitness - actions separated by risk tier, with named autonomous actions and a feature flag that can demote a send back to human review	0 1 2 3	
Governance readiness - every trial records retrieved policy, tool calls, final action, model and policy version, latency and trace ID, attributable to the agent and a named owner	0 1 2 3	

Resist the board grading itself

A private suite tuned until it passes is not a promotion board. Tick the safeguards that are actually in place.

- ☐ Hidden holdout tasks the build team has not seen
- ☐ Production samples nobody on the product team curated
- ☐ Flagged traces routed to a reviewer outside the product team
- ☐ A stale or contaminated case logged as a defect, not enjoyed as a result
- ☐ Evaluators structurally different from the generator - different model family, scaffold or evidence type

Choosing k without pretending k is magic

Repeated trials are a tool, not a ritual, and they are wasted if applied evenly. Tick what is true of your evaluation today; write next to each what would need to change.

- [ ] Stratified by action class and severity so a wrong order-status reply and a wrong refund are not averaged together
- [ ] Repeated trials reserved for tasks where variance actually matters, not run uniformly on every task
- [ ] A hidden holdout the build team has not seen, kept fresh release after release
- [ ] Reported as a confidence interval or honest range, not a lone decimal from a handful of trials
- [ ] Failed traces inspected, not only final scores
- [ ] Refreshed when the policy, the tools, the model or the user distribution changes

The promotion decision

State whether this action class may rise, must hold, or should drop a rung in the next release - and the single piece of evidence that would change the verdict.

Companion worksheet to **Essay 20 · You Cannot Benchmark a Coworker**, in the series **Architecting the AI Coworker**. · Dr Peter McCann Strain · Fill this in against one real agent, action class or vendor.

Series Companion + all 22 worksheets: **Release_v12/Series_Companion.pdf**