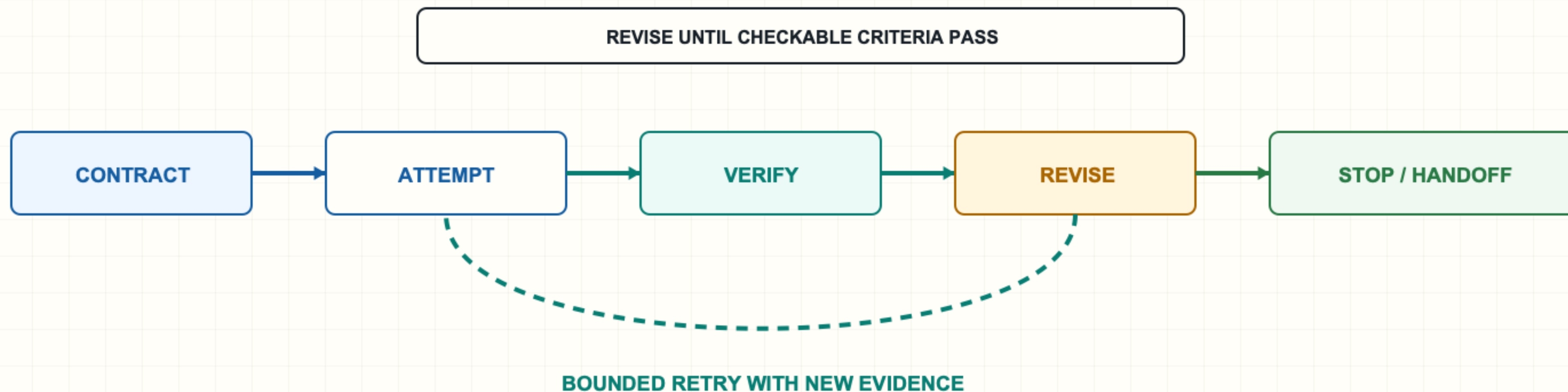


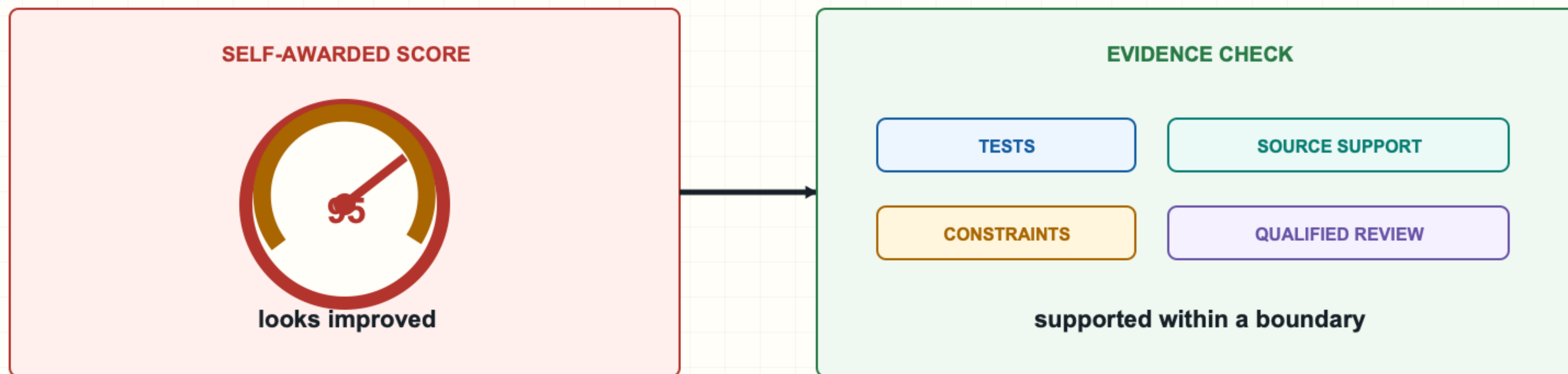
Engineering Agent Improvement Loops

From self-critique to verified revision.



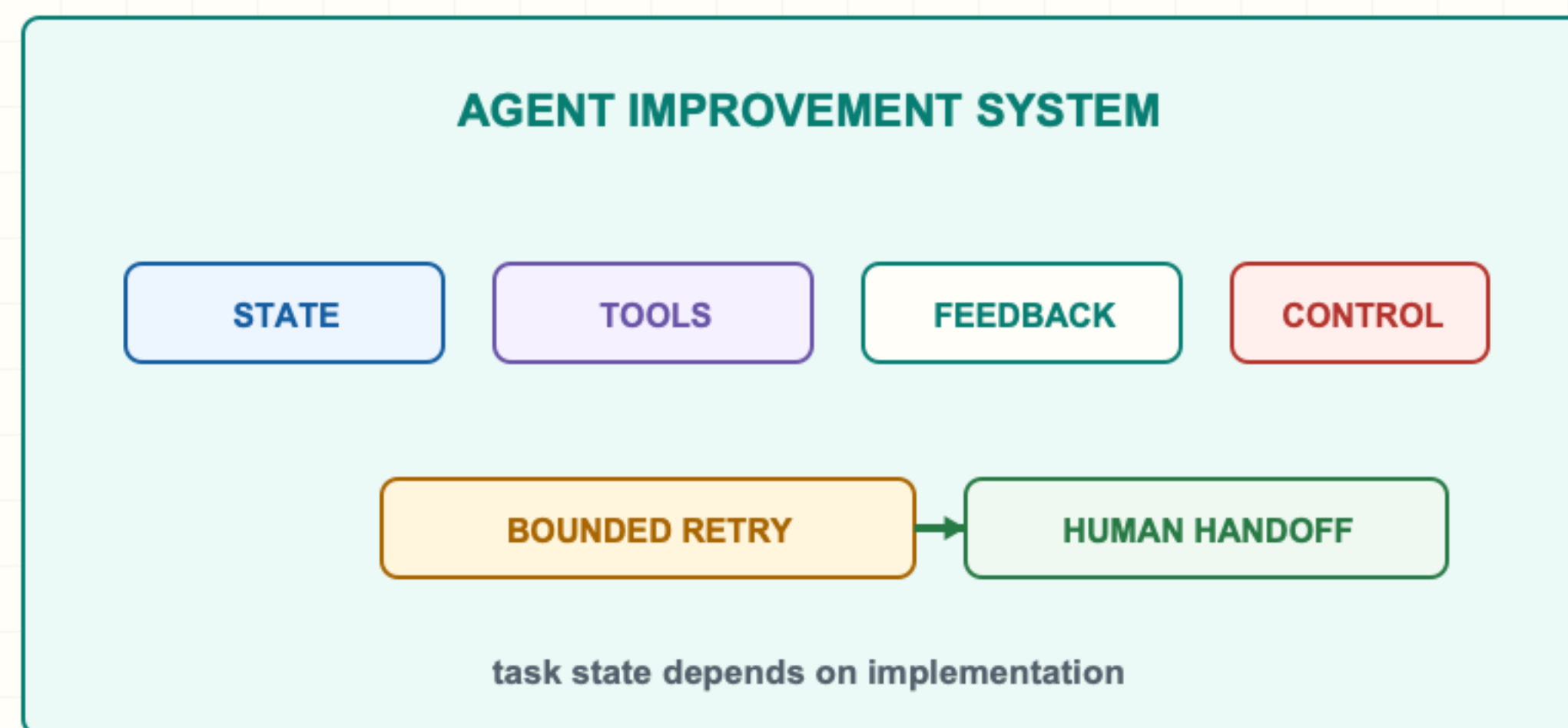
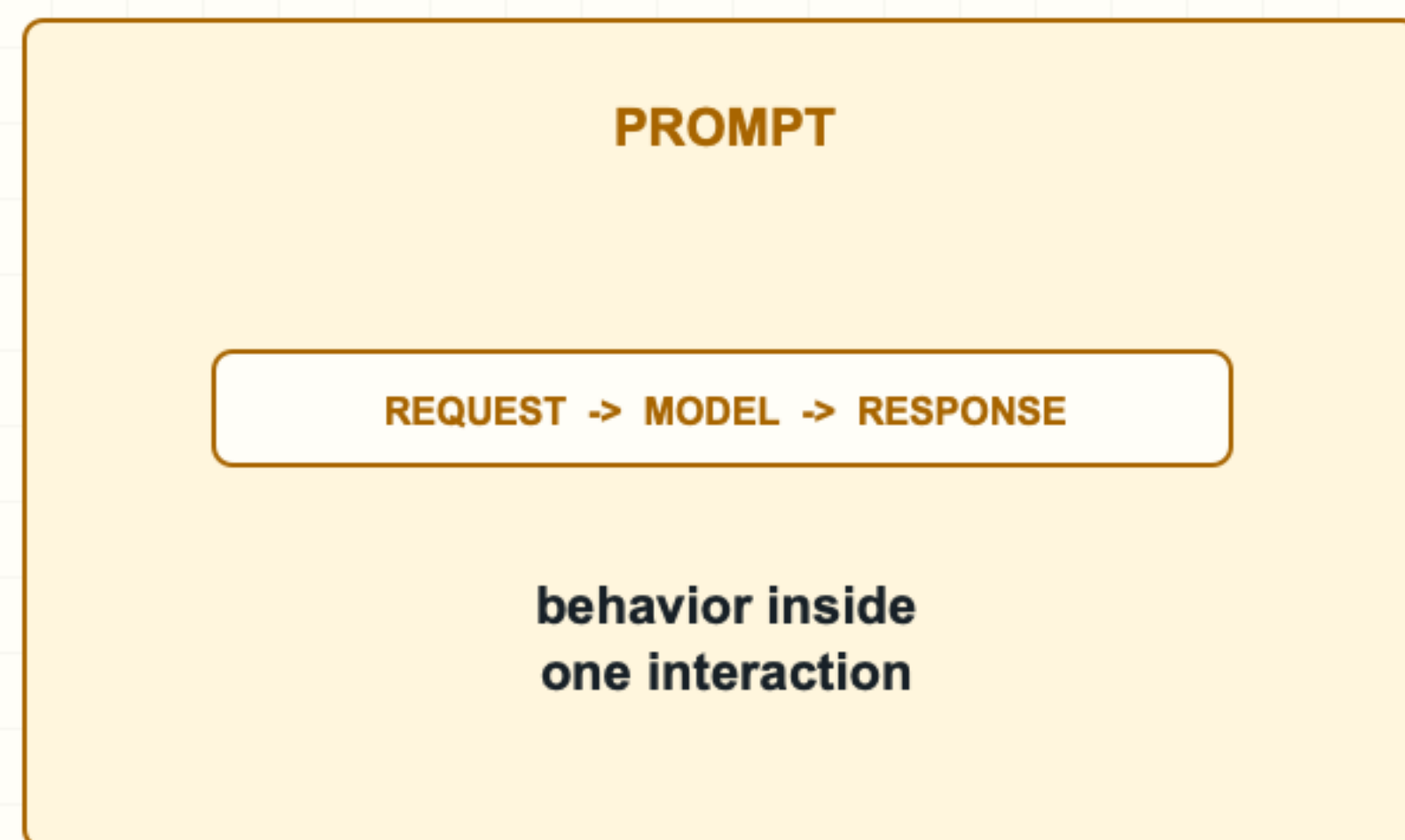
The Seductive 95-Point Prompt

Keep the loop. Reject the self-awarded score as proof.



A Prompt Is Not Yet an Agent

An improvement system adds state, tools, feedback, control, and handoff.



Four Research Patterns

Define the Opportunity and Limit

Refinement improves some tasks; external feedback changes reliability.

SELF-REFINE

Feedback: same model

Can improve some generation tasks; errors persist.

REFLEXION

Feedback: task result + memory

Carries textual lessons into later trials.

CRITIC

Feedback: external tools

Grounds revision in observable tool feedback.

INTRINSIC CORRECTION LIMIT

Feedback: none external

Can fail or degrade studied reasoning.

Use the Four-Question Gate

Specification, verification, repairability, and boundedness come before iteration.

01

SPECIFICATION

Can success be stated?

02

VERIFICATION

Can quality be checked?

03

REPAIRABILITY

Can defects be repaired safely?

04

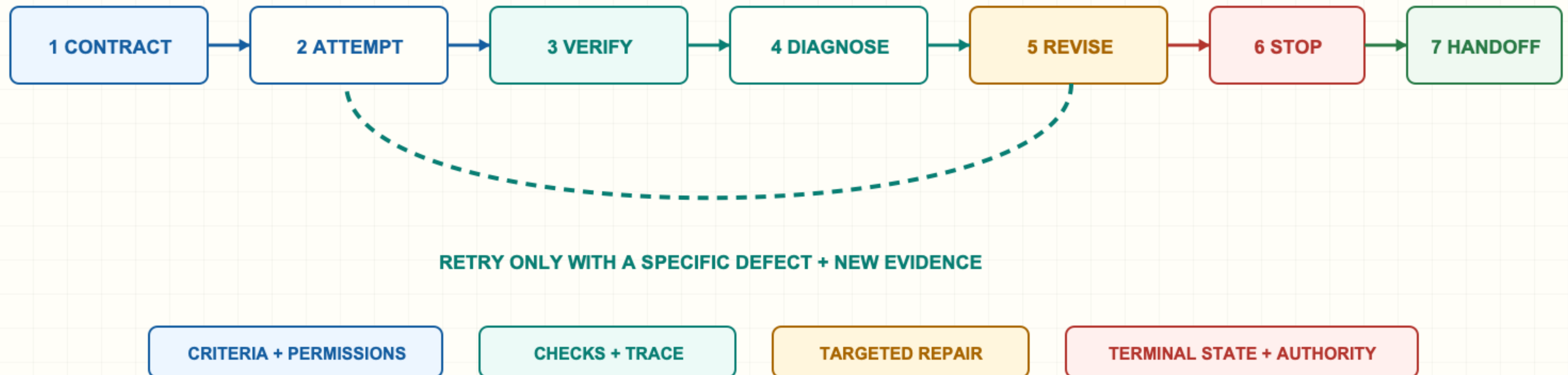
BOUNDEDNESS

Can cost, data, and authority be limited?

FOUR YES ANSWERS -> BOUNDED LOOP

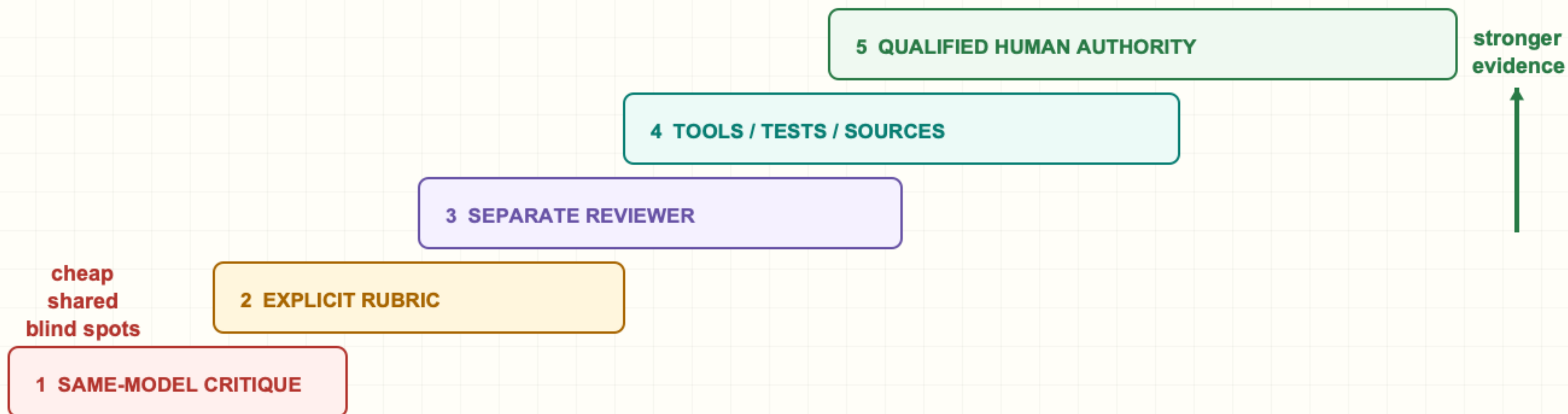
Build the Seven-Step Loop

Contract, attempt, verify, diagnose, revise, stop, and hand off.



Climb the Verifier Ladder

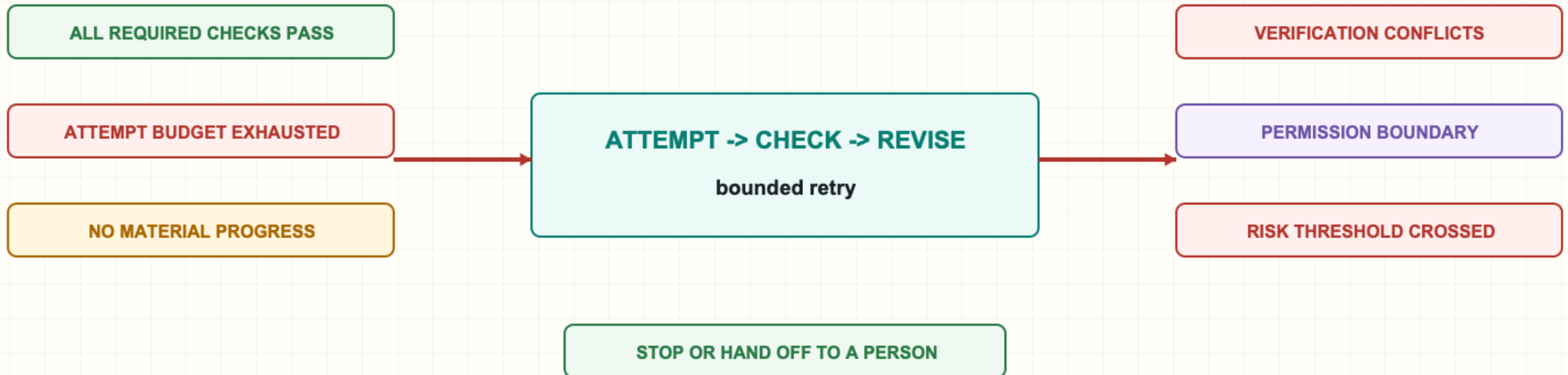
Independence can reduce correlated error; authority remains separate.



Independence may reduce correlated error. It does not create ground truth.

Stop Before the Loop Becomes the Problem

Every retry needs new evidence, a budget, and an escalation path.



Match the Loop to the Situation

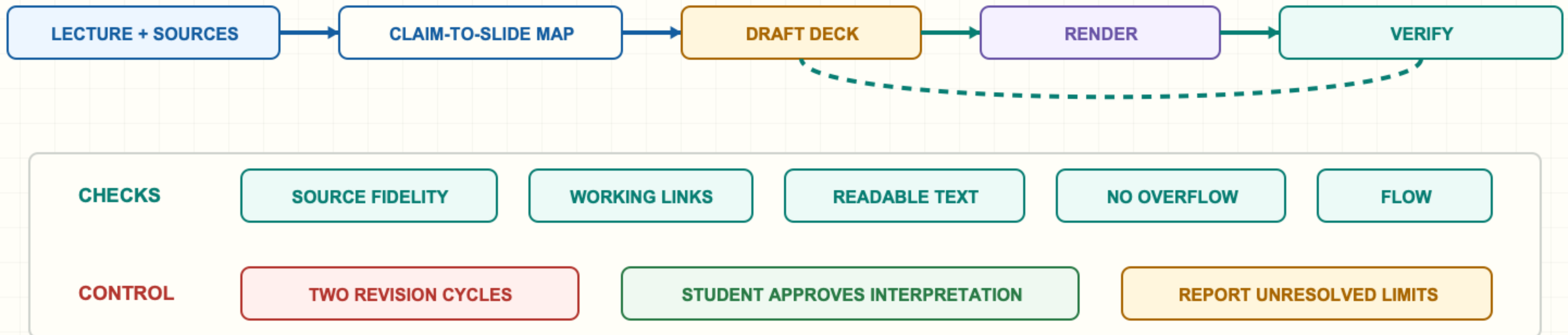
Different defects require different evidence and different human roles.

SITUATION	PRIMARY VERIFIER	AGENT ROLE
CODE	tests + review	OFTEN STRONG
RESEARCH	primary sources + interpretation	CONDITIONAL
SLIDES	source map + render checks	OFTEN STRONG
WRITING	instructor rubric + judgment	CONDITIONAL
CREATIVITY	creator / audience preference	HUMAN-LED
HIGH STAKES	domain evidence + authority	NARROW ROLE

Easy to score is not the same as valuable to optimize.

Worked Case: Lecture to Eight Slides

Fidelity, citations, legibility, and layout become observable checks.



Compare Three Conditions

Single pass, self-score to 95, and an evidence-verified loop.

A SINGLE PASS

ONE PROMPT

ONE SUBMISSION

B SELF-SCORE

REVISE TO 95

RECORD SCORES

C VERIFIED LOOP

CHECKS + EVIDENCE

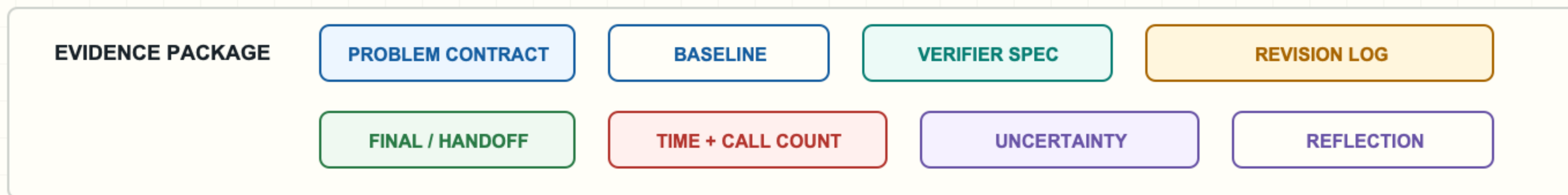
TWO-CYCLE LIMIT

COMPARE CORRECTNESS | COVERAGE | UNSUPPORTED CLAIMS | NEW DEFECTS | TIME | CALLS | PREFERENCE

Use prepared rubrics, starter artifacts, sources, checklists, and logs.

Submit Evidence, Not Just Polish

Grade the contract, verifier, revisions, stopping, evidence, and reflection.



The polished artifact is evidence, but it is not the whole grade.

The Closing Design Checklist

Loop on evidence, not confidence.

