TEACHER-FACING FRIDAY DESIGN

Reflection Loops: Critique, Revision, and Better Reasoning

CAS Human-AI Collaboration | Module 1 | Unit 6

LITERACY	COLLABORATION MOVE	CONTACT TIME	CHALLENGE
Interaction, secondarily Mechanism	**Reflection**	**365 minutes**	**Reflective Improvement Workflow**

Position in Module

- **Unit:** 6
- **Topic:** Reflection Loops: Critique, Revision, and Better Reasoning
- **Literacy perspective:** Interaction, secondarily Mechanism
- **Collaboration level:** Reflection
- **Fundamentals focus:** Generative AI Fundamentals
- **Why this Friday matters:** Participants learn that asking for improvement is not yet a reflection process. They learn to separate construction, judging, revision, and verification; define explicit conditions of satisfaction; and test whether critique produces evidence-backed improvement rather than confident churn.

Literacy Perspective

- **Primary literacy:** Interaction
- **Secondary literacy, if any:** Mechanism
- **What this literacy explains:** Interaction Literacy explains how evaluation criteria, role framing, artefact presentation, and feedback shape critique and revision. Mechanism Literacy explains why model judgments remain probabilistic and can vary with order, wording, verbosity, or contextual leakage.
- **What participants should manipulate, observe, and interpret:** Participants keep the original task and baseline artefact stable while adding an explicit, separated critique-and-revision loop. They observe defect detection, criterion coverage, judge stability, revision traceability, new errors, and net improvement, then interpret where model judgment helps and where external evidence or human review remains necessary.

Collaboration Level

- **Level relation:** Reflection
- **Collaboration move being learned:** Participants move from accepting or informally polishing an output to reflecting through explicit criteria, independent critique, reasoned revision decisions, and verification.

- **What participants should be able to do differently after this Friday:** They should be able to formulate conditions of satisfaction, separate builder and judge contexts, test a judge for instability, turn a critique into a selective revision plan, and demonstrate whether the revised artefact is actually better.

Generative AI Fundamentals

This Friday should introduce LLM-as-judge as a core pattern for reflective improvement. The goal is for participants to understand how generative AI can be used to evaluate artefacts, why this requires explicit conditions of satisfaction, and why judging should be separated from construction and revision.

- Explain that generative AI can be used not only to construct an artefact, but also to judge an artefact against explicit criteria. In this role, the model is asked to inspect, score, compare, or critique a result rather than produce the primary result directly.
- Introduce conditions of satisfaction as the central control mechanism for LLM-as-judge. A judge prompt is only as useful as the criteria it receives, such as factual support, completeness, consistency, audience fit, risk handling, test adequacy, or source traceability.
- Separate construction, improvement, and judging as different sessions or phases. The same undifferentiated chat should not be treated as a reliable producer, improver, and evaluator at once; clearer control comes from assigning each phase a distinct task, context, and artefact.
- Show how markdown documents or equivalent artefacts transfer state between sessions. A draft, criteria document, judge report, revision plan, and revised artefact make the loop inspectable and prevent the process from relying only on hidden conversation history.
- Make clear that LLM-as-judge remains generated judgment, not objective truth. It can miss defects, overstate certainty, invent problems, or accept weak reasoning, so its judgments should be compared with evidence, tests, sources, stakeholder feedback, or human review where the stakes require it.
- **Caveat to demonstrate:** an LLM judge can be biased by wording, order, verbosity, style, or knowledge of which output it produced. Running the same judgement with reordered candidates, anonymized outputs, or a separate judging session can reveal instability in the evaluation.
- **Emphasize the central learning point for Unit 6:** reflection loops improve generative work when construction, judging, and revision are separated by explicit artefacts and governed by clear conditions of satisfaction.

Recurring Agentic Coding Demonstration

Start with one fixed baseline implementation and its existing test result. In separate phases, provide explicit review criteria, ask an independent reviewer session or agent to produce a criterion-linked judge report, require the builder to accept, reject, or defer each finding with reasons, and run a targeted revision. Preserve the original diff, judge report, revision decision log, revised diff, and before/after tests. Repeat one judgment with anonymised or reordered evidence to reveal instability. The demonstration makes Interaction and Mechanism Literacies visible through criteria and probabilistic judging, while the tests show why a generated critique is evidence to examine rather than an objective verdict.

Shared Classroom Reflection Case

Use one common baseline artefact so that groups can compare judge behaviour and revision decisions against the same known issues.

Default case:

> **A fictional organisation is considering a six-month staff-mentoring pilot. A baseline AI-generated recommendation memo must advise a programme sponsor on whether and how to proceed, using the supplied brief and evidence notes.**

Prepare one fixed case pack containing:

- an aligned task brief with audience, decision need, constraints, and intended output
- a small evidence pack with participation estimates, workload constraints, one conflicting stakeholder view, and one explicit uncertainty
- one baseline memo containing a known mix of strengths and defects, such as an unsupported claim, omitted constraint, vague recommendation, inconsistent number, and unacknowledged uncertainty
- an instructor defect key that distinguishes seeded defects from arguable stylistic preferences
- a conditions-of-satisfaction template covering evidence support, completeness, internal consistency, audience fit, uncertainty handling, and actionability
- a judge-report template requiring criterion, finding, evidence location, severity, confidence, and proposed change
- a revision decision log with accept, reject, or defer plus reason
- two presentation variants for the judge-stability test, such as anonymised outputs in reversed order or reordered rubric criteria

All groups should begin from the same baseline memo. The target condition adds separated judging, revision planning, selective revision, and verification. Participants should not ask the judge to rewrite the memo directly, because that would hide the distinction between evaluation and construction.

Friday Activity Plan

Indicative contact time is 365 minutes, excluding breaks and lunch. Teachers may adjust durations while preserving baseline inspection, explicit criteria, judge testing, selective revision, verification, and transfer.

Time	Activity	Concrete output
40 min	Reflection is more than self-correction	Reflection-loop map and two failure predictions
45 min	Inspecting the common baseline	Baseline rubric scores and provisional defect list
45 min	From values to conditions of satisfaction	Shared rubric with anchors and evidence requirements
50 min	Testing the judge before trusting it	Two judge reports and stability comparison
60 min	Critique, decide, revise, verify	Revision decision log, revised memo, and verification record
55 min	Blinded before-and-after comparison	Comparison table, residual-risk note, and provisional reflection rule
70 min	Challenge kickoff and first reflective run	Challenge protocol, criteria set, prediction table, and first reflection artefact

Detailed activity blocks

1

40 min Reflection is more than self-correction

Content and manipulation. Distinguish construction, judging, revision, and verification. Introduce conditions of satisfaction and the limits of generated judgment.

Method and expected observation. Contrasted walkthrough of vague “improve this” interaction and separated reflection loop; participants predict where each can fail.

Concrete output. Reflection-loop map and two failure predictions

Tools and materials. Slides, contrasted traces, loop cards

2

45 min Inspecting the common baseline

Content and manipulation. Review the fixed mentoring memo against the task brief and evidence before using an AI judge. Capture individual judgments to avoid immediate anchoring.

Method and expected observation. Silent annotation followed by small-group calibration; expected observation is that even humans interpret broad criteria differently.

Concrete output. Baseline rubric scores and provisional defect list

Tools and materials. Case pack, baseline memo, annotation sheet

3

45 min

From values to conditions of satisfaction

Content and manipulation. Turn broad qualities such as “good reasoning” or “professional” into anchored, inspectable criteria. Define what evidence would count.

Method and expected observation. Criteria mini-lab and cross-group test on one paragraph.

Concrete output. Shared rubric with anchors and evidence requirements

Tools and materials. Criteria cards, rubric template, sample paragraph

4

50 min

Testing the judge before trusting it

Content and manipulation. Run the same evaluation in separated judge sessions using anonymised or reordered presentation. Compare findings, severity, and confidence.

Method and expected observation. Controlled judge-stability experiment; participants identify stable findings, order effects, false positives, and missed seeded defects.

Concrete output. Two judge reports and stability comparison

Tools and materials. Chat interface or prepared judge outputs, presentation variants

5

60 min

Critique, decide, revise, verify

Content and manipulation. Use one judge report to create an accept/reject/defer decision log, then revise only accepted findings and check the result against sources and criteria.

Method and expected observation. Role-separated work with judge, revision owner, and verifier. Expected observation is that selective revision is more controllable than automatic rewriting.

Concrete output. Revision decision log, revised memo, and verification record

Tools and materials. Judge report, markdown workspace, evidence pack

6

55 min

Blinded before-and-after comparison

Content and manipulation. Compare baseline and revised artefacts without revealing which is which. Check defect correction, new defects, criterion scores, and residual uncertainty.

Method and expected observation. Peer review followed by evidence-based interpretation; each group states one improvement and one claim the evidence cannot support.

Concrete output. Comparison table, residual-risk note, and provisional reflection rule

Tools and materials. Blinded artefacts, rubric, shared board

7

70 min **Challenge kickoff and first reflective run**

Content and manipulation. Select an artefact where critique matters, define baseline and target loop, draft criteria, and capture a first baseline assessment or judge report.

Method and expected observation. Individual or paired coaching with feasibility and judge-independence check.

Concrete output. Challenge protocol, criteria set, prediction table, and first reflection artefact

Tools and materials. Challenge template, judge-report template, revision log

Experiment Focus

- **Experiment focus:** direct output versus a separated critique, revision-decision, and verification loop governed by explicit conditions of satisfaction.
- **Primary manipulated condition:** presence of the explicit reflection loop. The target should add criteria, an independent judging phase, a documented revision decision, selective revision, and verification.
- **What remains constant across comparison:** Keep the original task, source material, baseline artefact, intended audience, and evaluation criteria stable. The target should revise the same baseline rather than regenerate a different task from scratch.
- **Judge-stability check:** Repeat at least one evaluation with anonymised or reordered presentation, or in a separate session, to test whether findings remain stable.
- **Prediction requirement:** Before judging, predict which defect categories should improve and which risks, including new errors or style bias, may remain.

Observation Focus

- **Criterion coverage:** Does the judge address every condition of satisfaction or concentrate on easy stylistic issues?
- **Defect detection:** Which seeded or independently verified problems are found, missed, or invented?
- **Judge stability:** Do findings, severities, and rankings change with order, anonymity, wording, or session context?
- **Revision traceability:** Can each material change be connected to a finding and an explicit accept, reject, or defer decision?
- **Net improvement:** Which defects were corrected, which remain, and which new defects appeared?
- **External validity:** Which claims require source checks, tests, stakeholder feedback, or human expertise beyond the judge report?
- **Evidence participants should capture:** baseline artefact; criteria and anchors; judge prompts and reports; presentation variants; revision decision log; revised artefact; before/after rubric scores; verification evidence; and unexpected results.
- **Interpretation participants should be prepared to make:** Explain which parts of the loop improved the artefact, where judge instability or false confidence limited control, and why a higher judge score alone is not sufficient evidence of better reasoning.

Challenge Kickoff Deliverables

Participants must leave Friday with:

- a selected baseline artefact whose quality matters and can be evaluated without exposing confidential material
- a stable task brief, source basis, intended audience, and intended outcome
- explicit conditions of satisfaction with observable anchors
- a target reflection loop separating construction, judging, revision decisions, selective revision, and verification
- a prediction table identifying expected improvements and judge-related caveats
- a protocol draft covering the baseline, manipulated process condition, constants, judge-independence strategy, evaluation criteria, and verification method
- a judge-report and revision-decision template tailored to the selected artefact
- **at least one first captured artefact:** baseline rubric, judge report, stability check, revision decision, or peer feasibility review
- a plan for evidence outside the LLM judge when factual, technical, or stakeholder claims require it

The kickoff should already reflect the challenge-focus logic:

- **Difficulty:** Direct generation leaves weaknesses, blind spots, and unsupported claims unexamined.
- **Approach:** Define explicit conditions of satisfaction, separate builder and judge contexts, test judge stability, document revision decisions, and compare the verified revision with the same baseline.
- **Artefacts:** protocol document; prediction table; baseline artefact; anchored criteria; judge reports; stability comparison; revision decision log; revised artefact; external verification evidence; before/after run log; concise interpretation; and one practical reflection rule.

Teacher Setup

Mandatory:

- one shared aligned brief, evidence pack, and baseline artefact with a known mix of strengths and seeded defects
- an instructor defect key that distinguishes verifiable problems from preferences
- conditions-of-satisfaction rubric with observable anchors
- judge-report, presentation-variant, revision-decision, verification, comparison, prediction, and protocol templates
- a method for separated judge sessions and anonymised or reordered evaluation
- demonstration setup for the recurring agentic coding case, if used
- prepared judge reports and paper-based comparison fallback so the reflection experiment does not depend on live model access

Configurable:

- case domain and artefact type
- exact model or interface used for construction, judging, and revision
- whether the same model is used in separate sessions or a different judge model is available
- **judge-stability manipulation:** order, anonymity, rubric order, or independent repeated judgment
- work mode and role allocation
- number and type of external verification sources, tests, or reviewers
- scaffolding depth and peer-feedback format