Human-Centered AI Deployment Readiness Protocol

Engineering-handoff profile · v0.1-rc.2 · Yejun Tak · September 8, 2026

Purpose, scope and status

This protocol helps human evaluators examine whether an AI-generated interface prototype contains enough requirement, interaction-state and recovery evidence for engineering handoff. It separates an evaluator's readiness judgment from the evidence supporting that judgment. Version 0.1-rc.2 is a proposed method for testing and refinement. Its example is synthetic. No external validation, measured effectiveness, production certification or organizational adoption is claimed.

Deployment readiness describes the wider program. This initial engineering-handoff profile covers prototype specifications and walkthroughs. It does not establish production readiness or replace functional, accessibility, security, privacy, model-performance or regulatory testing. Record whether AI generated the interface, operates at runtime, both, or neither. These are different populations. The included fictional case has no runtime AI.

Intended users are UX evaluators, product designers, engineering leads and researchers reviewing a frozen prototype and task brief. A project owner must define the relevant stage and requirements before evaluation. Visual polish is not itself a defect. A difference between perceived readiness and criterion status is an observation to investigate, not proof that visual fidelity caused it.

Start with the separate evaluator packet described in Templates/evaluator-packet.md, retain a separate reference key, lock the evaluator's findings and judgment, then use Evaluation-Template.xlsx to calculate results. Worked-Example contains a fictional artifact specification, reference records and completed calculations. Do not show the reference key to evaluators before their records are locked.

1. Freeze the assessment

Record session ID, artifact/version, date, intended user and task, assessment stage, evaluator, decision owner, time limit, environment, exclusions and presentation conditions. Specify whether the purpose is artifact assessment, evaluator assessment or both. Record model/tool provenance when known; do not guess how an artifact was generated.

Freeze mandatory requirements, applicable recovery scenarios, acceptance checks, severity rules and the gate before the evaluator sees the artifact. Every exclusion needs a reason. A changed criterion creates a new dated version. Keep the old record.

The provisional v0.1 handoff gate requires every mandatory requirement and applicable recovery scenario to be both specified and walkthrough verified, no unresolved critical defect, no unassessed mandatory item, and a complete evidence record. All conditions must be satisfied. This conservative project rule is proposed here, not established by NIST or empirically validated. An eligible result goes to the designated owner for handoff review.

A specified behavior describes trigger, action, resulting state and preserved or changed data. Walkthrough verified means a reviewer has followed the defined path in the frozen artifact and found that it satisfies the acceptance check. Implemented and runtime tested are separate evidence levels. A drawn button alone cannot verify recovery behavior.

2. Roles, reference and severity

The artifact owner supplies the version and requirements. A reference reviewer prepares and checks the criterion record and defect key. An evaluator independently examines the artifact. An adjudicator resolves finding matches and disputes. The decision owner records the final disposition. Disclose role overlap. For an independent evaluation, the evaluator should neither have developed the artifact nor seen its reference key.

A seeded benchmark uses a frozen set of deliberately included defects. A second reviewer should check observability and match rules. A real prototype uses a documented, adjudicated reference set that may still be incomplete. Recall against that set is not the proportion of every actual defect found.

Keep the answer key separate until judgments are locked. New genuine defects affect the artifact decision where relevant, but must not silently change the denominator of the frozen recall measure. Report them separately and issue a versioned correction or exploratory recalculation if warranted.

Critical: an essential failure with a serious consequence in the scoped task, such as omitting a required confirmation for a consequential action. Major: a required task is prevented or a materially wrong but recoverable outcome is likely. Minor: limited friction or ambiguity that does not invalidate a mandatory task. Document consequence and context, not just the label. Preserve disagreements.

Use categories such as requirements, system status, recovery and user control. For runtime-AI interfaces, document applicable correction, uncertainty and escalation requirements while treating model performance as separate evidence outside this profile.

3. Run and lock the evaluation

1. Supply only the evaluator-facing brief, requirements-brief.csv, recovery-brief.csv, artifact, and Evaluator-Scorecard.pdf. Do not supply the reference-location field, completed outcome matrices, answer key, completed workbook or entire repository to a blinded evaluator. A public training case cannot serve as an unseen test case for someone who has studied it. Withhold reference outcomes, seeded-defect keys and other evaluators' findings. Use the same instructions and time limit within a comparison condition.

2. Complete assigned tasks, including failure and recovery paths. Record each finding with ID, requirement/state ID, reproduction steps, observed and expected behavior, proposed severity, evidence location and uncertainty. Preserve raw findings. Do not merge or edit an evaluator's words after locking.

3. Before revealing the key, ask: "What percentage of the predefined reference defects do you estimate your findings correctly identify? Enter 0-100." Do not reveal the number of reference defects. This is an expected-recall estimate, not the probability of having found every defect.

4. Ask separately: "Under the supplied engineering-handoff criterion, is this artifact Ready, Not ready, or Unable to assess?" Require a reason. Record elapsed minutes and the locking timestamp with timezone. The evaluator's judgment is distinct from the final criterion determination.

5. Reveal the key and adjudicate each finding as matched, duplicate, unsupported, novel genuine, or unresolved. A match must identify a specific deficiency. Each reference defect counts at most once per evaluator. Record IDs, adjudicator, rationale and unresolved disagreement. Do not count vague suggestions as detected defects.

6. Reconcile workbook counts to the retained records. Determine criterion status from requirement, recovery, critical-issue and evidence records. Calculate the applicable measures. The owner records disposition, conditions and the next review trigger.

4. Define the measurements

Reference-set defect recall. Unique correctly detected reference defects / eligible reference defects. Report both counts and percent. A zero or unknown denominator is N/A. Preserve duplicate, unsupported, novel and unresolved findings separately. A novel genuine finding is not a frozen-reference true positive.

False-ready acceptance. Among criterion-nonready assessment instances with an observed judgment, divide Ready judgments by all observed judgments in those instances. One evaluator assessing one artifact is one instance. Unable to assess is an observed abstention included in that denominator; also report its rate. Missing judgments are excluded and counted separately. Unknown criterion status is excluded and separately counted. A gate failure caused by missing required evidence is criterion-nonready, not unknown.

Expected-recall gap. Expected recall (%) minus observed reference-set recall (%), in percentage points. Positive means overestimation; negative means underestimation. An 80% estimate and 62.5% recall yield +17.5 pp. Do not substitute confidence in success or the probability of finding every defect. Report per-session gaps and their distribution before considering aggregation.

Handoff recovery coverage. Applicable required recovery scenarios that are specified and walkthrough verified / all applicable required recovery scenarios. Unassessed required scenarios stay in the denominator. Zero applicable scenarios yields N/A with documented applicability reasoning. Preserve implementation and runtime-test status separately.

Requirements completeness recognition. Correctly identified predefined requirement omissions / all predefined reference omissions. This measures the evaluator. Separately report artifact requirements coverage: mandatory requirements specified and walkthrough verified / all mandatory requirements. The two measures have different numerators, denominators and interpretations.

No weighted total or universal pass percentage is used. Recall and expectation concern the evaluator. Coverage and critical issues concern the artifact. Strong recall cannot make a deficient artifact eligible. A false-ready rate is not the fraction of Ready judgments that were wrong, which uses a different denominator.

Companion measures and interpretation

False-ready acceptance must be accompanied by abstention, missingness and decision coverage. Decision coverage on criterion-nonready instances is (Ready + Not ready judgments) / all observed judgments on those instances. Decisive false-ready acceptance is Ready / (Ready + Not ready) on criterion-nonready instances. When all judgments abstain, the original false-ready rate is zero, decision coverage is zero, and decisive false-ready acceptance is N/A. This is not evidence of useful discrimination.

Include independently adjudicated criterion-ready control cases when studying effectiveness. Report Not ready judgments / all observed judgments on criterion-ready cases as the false hold rate, alongside the ready-case abstention rate and missing count. These complements are descriptive diagnostics, not new constructs or an aggregate score. Do not claim superiority by reducing acceptance through universal refusal or abstention.

The expected-recall gap is a session-level signed discrepancy. It is not a validated general measure of metacognitive ability or probability calibration. Report absolute discrepancies as an additional descriptive statistic if needed; signed gaps can cancel across sessions. Pre-specify aggregation and retain per-session values. Recall can increase through indiscriminate reporting, so also retain unsupported, duplicate and unresolved finding counts, adjudication burden and elapsed time.

5. Decide, report and reproduce

Hold for remediation: a documented failure or unresolved critical issue prevents the gate from being met. Insufficient evidence: mandatory checks or required records remain incomplete without an established qualifying pass. Eligible for handoff review: all gate conditions are met. The named owner must still record approval or refusal and date. The workbook reports gate eligibility, not automatic handoff approval.

The workbook uses manual, adjudicated counts for one session and a separate batch table for criterion-conditioned judgments. Reconcile every count to CSV records. Blank means missing; zero means assessed and none. Enter 0.80 (80%) for an expected recall of 80%, not 80. Counts must be whole and nonnegative. Every numerator must fit its denominator. Detected omissions must be a subset of detected reference defects, and detected non-omissions cannot exceed the non-omission reference set. In the gate inputs, verified checks + documented failed checks + unassessed checks must equal all mandatory requirement and applicable recovery rows. Count each row in exactly one status. Requirement and recovery rows are distinct checks even when they share a defect. Open critical defects are a separate count. Workbook B36 records documented failed checks; a missing required count yields Not evaluated.

Retain the frozen brief; requirement/recovery matrix; reference key; raw findings; pre-key judgment; adjudication; decision and revision record. The supplied CSV headers define stable record IDs and evidence locations. Copy the template for each new assessment. The batch workbook supports 20 records; for larger studies use exported records and an explicitly reviewed aggregation script rather than silently exceeding its range.

Report artifact population, criterion/version, actual sample, roles, relationships, missing records, abstentions, reference uncertainty and material deviations. For batch rates, avoid pooling different criteria without stratification. Repeated judgments on an artifact or by an evaluator are not statistically independent observations. Do not present this descriptive workbook as an inferential analysis.

Re-evaluate changes against affected requirements and states. Preserve the prior version and decision. A remediation improvement on one artifact does not isolate a causal effect of this protocol. Publish only original or permitted materials; omit private product data and participant identifiers.

6. Worked example and next validation

The fictional service-request case has 10 mandatory requirements, 6 applicable recovery scenarios and 8 stipulated reference defects. The synthetic evaluator detects D01, D02, D04, D05 and D07. Three requirements are omitted (R05, R07, R10), of which only R07 is recognized. Expected recall is set to 80%. One critical issue remains open.

These inputs yield 5/8 = 62.5% reference recall; 80% - 62.5% = +17.5 pp expected-recall gap; 3/6 = 50.0% recovery coverage; 1/3 = 33.3% omission recognition; and 7/10 = 70.0% artifact requirements coverage. The artifact is held for remediation. Four fictional batch records contain three Ready judgments and one Not ready judgment on criterion-nonready artifacts, giving 3/4 = 75.0% false-ready acceptance.

The example is a constructed, specification-based exercise. It has not been tested with participants, independently validated as a stimulus, or run on an employer product. The five stipulated minor clarity issues do not invalidate their mandatory requirements in this example. A real reference reviewer may disagree and should document the change before evaluation.

First conduct an external desk review of instructions, observability, match rules and spreadsheet use. Log requested changes and disagreements. Before covered participant recruitment or data collection, obtain the institution's applicable research determination. The pilot kit supports a bounded feasibility exercise; it does not claim institutional approval.

A proposed 6-12 practitioner feasibility exercise may examine instruction clarity, duration and adjudication consistency. This is a feasibility target, not a power calculation. A comparative study additionally needs independently reviewed matched cases, prespecified assignment and analysis, and carryover controls. This release supplies one instructional case, not a validated matched pair. Report negative findings and limitations in v0.2.

7. Research context and provenance

The AI RMF provides context for explicit requirements, independent assessment, documented measurement methods and evaluation-method validation [1]. This is a selective conceptual relationship, not conformance certification. The protocol's definitions and provisional handoff gate are application-specific proposals.

The TEVV-Athlon initial draft addresses evaluation design and evidence synthesis [2]. Tak's separate September 6, 2026 comment proposes documenting relevant presentation conditions and examining meaningful divergence between human judgments and independently assessed evidence. The retained sent-email record documents transmission; it does not establish receipt, NIST adoption or endorsement. This protocol was developed as a separate artifact and should not be described as an attachment to that earlier submission.

[1] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, January 2023. Core tables, especially MAP 1.6 and MEASURE 1.3, 2.1, 2.3 and 2.13. https://doi.org/10.6028/NIST.AI.100-1

[2] Phillips, P. J., et al. The TEVV-Athlon Framework for Evaluating AI Systems. NIST AI 200-2 ipd, August 2026. Section 2.4 and Appendix E, Table 7. https://doi.org/10.6028/NIST.AI.200-2.ipd

Version: 0.1-rc.2, September 8, 2026. Author attribution: Yejun Tak. Development assisted by OpenAI Codex for drafting, packaging and arithmetic checks. Author technical review and external validation remain

separate steps. This correction release records AI-assisted literature comparison and technical checks, not an independent expert review or empirical validation. No DOI has been assigned at package preparation.

 Corrections should identify version, section or stable record ID and explain the proposed change.

8. Prior work and contribution boundary

This release is an operational synthesis for a narrow engineering-handoff assessment. It does not claim to invent usability inspection, coverage, classification error rates, confidence-performance discrepancy or human-centered AI readiness. Its proposed contribution is the combination of a frozen stage criterion, a judgment locked before reference disclosure, separate evaluator and artifact measures, and retained evidence records in one reusable workflow. Whether that combination improves decisions over a strong existing checklist remains untested.

Nielsen's usability heuristics already address system status, user control, error prevention and recovery [3]. Amershi and colleagues provide 18 human-AI interaction guidelines evaluated with design practitioners [4]. Google's People + AI Guidebook also provides practical treatment of failures, uncertainty and fallback paths [5]. These works ground the content of checks; they are not evidence that this protocol's procedure is new or effective.

The ML Test Score supplies a production-readiness rubric for ML systems [6]. HINT operationalizes predeployment evaluation of AI features with human participants [7]. Lee's 2026 readiness taxonomy covers observable human-AI behavior, calibration and harm [8]. The present profile instead focuses on the evaluator's engineering-handoff judgment about a frozen interface specification. This narrower unit of analysis is a proposed distinction, not a priority claim.

Expected-recall discrepancy draws on established metacognition measurement distinctions [9]. Prior prototype-fidelity research includes both effects under particular conditions and null findings for perceived usability [10, 11]. The synthetic text specification in this release does not manipulate visual polish, establish an AI-specific effect or show that fidelity causes false readiness.

Evaluator disagreement is a known problem in usability assessment [12]. Independent reference preparation, documented match rules and agreement analysis are therefore essential validation work. A completed workbook proves only that specified calculations ran, not that reference defects, severity or gate thresholds are valid.

[3] Nielsen, J. 10 Usability Heuristics for User Interface Design. https://www.nngroup.com/articles/ten-usability-heuristics/

[4] Amershi, S., et al. (2019). Guidelines for Human-AI Interaction. CHI. https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[5] Google PAIR. People + AI Guidebook, Errors + Graceful Failure. https://pair.withgoogle.com/chapter/errors-failing/

[6] Breck, E., et al. (2017). The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. IEEE Big Data. https://research.google/pubs/the-ml-test-score-a-rubric-for-ml-production-readiness-and-technical-debt-reduction/

[7] Chen, C., Schnabel, T., Nushi, B., and Amershi, S. (2022). HINT: Integration Testing for AI-based Features with Humans in the Loop. IUI.

https://www.microsoft.com/en-us/research/publication/hint-integration-testing-for-ai-based-features-with-humans-in-the-loop/

[8] Lee, M. H. (2026). From Accuracy to Readiness: Metrics and Benchmarks for Human-AI Decision-Making. CHI Extended Abstracts. https://arxiv.org/abs/2603.18895

[9] Fleming, S. M., and Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443. https://doi.org/10.3389/fnhum.2014.00443

[10] Sauer, J., and Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests. Applied Ergonomics, 40(4), 670-677. https://pubmed.ncbi.nlm.nih.gov/18691696/

[11] Wiklund, M. E., Thurrott, C., and Dumas, J. S. (1992). Does the Fidelity of Software Prototypes Affect the Perception of Usability? https://doi.org/10.1177/154193129203600429

[12] Hertzum, M., and Jacobsen, N. E. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. International Journal of Human-Computer Interaction, 13(4), 421-443. https://doi.org/10.1207/S15327590IJHC1304_05