

GOVERNMENT PILOT EVALUATION CHARTER

The system is ready. Put it through a controlled pilot.

Deep Fathom has built the complete Level 2 model, source-linked evidence, repeatable checks, guided remediation, and AI-assisted review. This charter gives Government a controlled way to vet that system: same cases, independent answers, measured errors, and no effect on any contractor.

THE CONTROL

No experimental output affects certification, SPRS, award, affirmation, enforcement, supplier status, or procurement preference.

01 / THE QUESTION

Where does the working system reduce effort while quality and safety hold?

Run ordinary and assisted review on the same cases. Count missed evidence, unsupported conclusions, conflicts found, corrections, reviewer time, and disagreement.

02 / WHO CONTROLS THE ANSWER

The vendor does not grade its own system.

Write the rules first. Independent reviewers establish the answers without vendor output. Use the same inputs. Keep disagreements. Publish failures and wins.

IMMUTABLE SAFEGUARDS

The experiment cannot make a CMMC decision.

It can show where a method helps or fails. Its records stay outside certification, SPRS, awards, affirmations, enforcement, and supplier status.

- 01

No program effect
Experiment records remain isolated from certification, SPRS, award, affirmation, enforcement, supplier-status, and procurement-preference systems and decisions.
- 02

Government owner
A named sponsor decides what the test may use, what it may not affect, who is responsible, when to stop, what can be disclosed, and which later decision the results will inform.
- 03

Independent answers
Qualified Government-designated reviewers establish the reference findings without seeing vendor output. Their disagreements and final resolution remain in the record.
- 04

Rules written first
Record the cases, sources, measures, pass and stop thresholds, software and model versions, important settings, and correction rules before scoring starts.
- 05

People make findings
Software may organize, map, compare, detect, and prioritize. Qualified reviewers decide what the evidence supports. Authorized officials retain every Government decision.
- 06

Corrections stay visible
A correction says what was wrong, what changed, who approved the change, and which dependent findings must be checked again. A silent rerun is not a correction.

REQUIRED PARTICIPATION

Problem owner Defines the decision and mission value.	Program and method owner Sets the CMMC and assessment boundary.	Reference reviewers Establish and adjudicate findings.
Data and security owners Authorize every input and environment.	Acquisition authority Determines any lawful funded path.	Independent methods Receive the same frozen test where feasible.

The vendor does not control the answer. It does not authorize the data, establish the reference finding, or decide whether its own result passes.

PILOT PROTOCOL

Write the rules. Compare the same cases. Publish what failed.

A later reviewer should be able to reproduce every scored run from the saved cases, versions, settings, outputs, and corrections.

01

Freeze the test

State the objective, users, permitted and prohibited uses, cases, inclusion rules, methods, versions, measures, success floors, stop thresholds, subgroup analysis, and correction policy. Register exceptions before the scored run.

Record: signed charter, protocol, version manifest, analysis plan.

02

Build the reference

Qualified reviewers assess versioned inputs independently, cite supporting sources, record confidence and uncertainty, and adjudicate disagreement without seeing method output.

Record: findings, citations, disagreement log, adjudication.

03

Run blind comparisons

Human and AI-assisted methods receive the same inputs where feasible. Output separates source, deterministic check, machine analysis, human finding, and authorized decision. Material method changes create a new run.

Record: attributable output, timing, labor, logs, errors.

04

Examine the errors

Review unsupported conclusions, missed conflicts, false negatives, disagreement, and corrections. State who could be harmed by relying on a wrong answer and where independent human review remains necessary.

Record: error type, severity, cause, and correction.

05

Publish the limits

Report labor, accuracy, safety, case-by-case results, exceptions to the protocol, and where the result does not apply. A successful experiment does not grant status to a supplier or product.

Record: results, limitations, and decision memo.

A passing average is not enough.

Advance only if burden improves while predeclared quality and safety floors hold across the relevant cases.

CASES AND MEASURES

Test difficult cases, not demo cases.

Synthetic inputs come first. Controlled representative inputs follow only with separate authority, protected handling, and a defined evaluation purpose.

01	Technical	Configuration, identity, logging, vulnerability, network, endpoint, and source-system evidence.	02	Procedural	Policies, interviews, demonstrations, judgment-dependent objectives, and implementation gaps.
03	Inherited	MSPs, ESPs, cloud services, shared responsibility, customer duties, and provider limitations.	04	Constrained	Operational technology, legacy systems, small suppliers, compensating conditions, and sparse telemetry.
05	Degraded	Incomplete, stale, contradictory, duplicated, mislabeled, and internally inconsistent packages.	06	Adversarial	Misleading artifacts, hidden instructions, poisoned context, suspicious sources, and crafted omissions.
07	Changed	Material change after disposition, new inheritance, altered scope, expired evidence, and dependent findings.	08	Ambiguous	Reasonable reviewer disagreement, uncertain responsibility, insufficient support, and appeal.

MEASURE THREE DIMENSIONS SEPARATELY

Burden Supplier evidence-production labor; reviewer labor; duplicated handling; elapsed assembly and update time; cost separated into security, evidence, and verification work.	Quality Coverage; source traceability; supported findings; missed conflicts; agreement and disagreement with adjudicated findings; reproducibility across repeated runs.	Safety Unsupported conclusions; false negatives; unsafe reliance; adversarial robustness; drift; correction and appeal behavior; subgroup performance.
--	--	--

Report distributions, not just averages. Separate results by case, evidence type, supplier context, and operating environment. The cases where a method fails may matter more than its mean score.

DATA, TRANSITION, AND NEXT ACTION

Use synthetic cases first. Approve real data separately.

The fact that software can process data does not give anyone permission to use it. Every expansion of data, environment, user, or purpose requires a new decision.

PHASE 1	SEPARATE AUTHORIZATION	NEW DECISION
Synthetic benchmark Versioned cases, independently established reference findings, blind comparison, and predeclared measures. No FCI, CUI, customer material, supplier evidence, or production credentials.	Controlled representative inputs Only in a Government-approved environment with authorized owners, minimum necessary data, access control, encryption, audit, retention, incident handling, disclosure review, and verified destruction.	Operational prototype Requires a new scope, authority, safety case, security boundary, acceptance tests, acquisition determination, and transition plan. Evaluation success does not create automatic authorization.

Default data rule	No vendor model training on Government or participant data. No secondary use. Least-privilege access. Attributable audit events. Exportable records. Approved retention and destruction. Final terms remain the responsibility of authorized Government officials and the selected instrument.
--------------------------	--

THE EVALUATION SHOULD LEAVE THE GOVERNMENT WITH

• an inspectable protocol, case corpus, and reference-finding record; • revised vendor-neutral evidence profiles and validation rules; • procurement-ready performance requirements and human-decision controls, not a self-certified product claim.	• burden, quality, safety, subgroup, failure, and limitation results; • security, data-handling, records, rights, export, and transition artifacts; and
--	---

THE FIRST 45 MINUTES

Choose the workflow, data, reference reviewers, measures, stop rules, and owners.

The meeting should end with named owners, a protocol workshop date, and the questions Government expects the pilot to answer.

SOURCES AND STATUS

Deep Fathom discussion draft, reviewed August 13, 2026. Public design: cmmc20x.com/evaluation. Evidence example: cmmc20x.com/evidence. Current CMMC policy boundary: [32 CFR part 170](#). Structured assessment concepts: [NIST OSCAL assessment layer](#). AI risk framing: [NIST AI RMF](#). This charter is not Government policy, an authorized method, a solicitation, or an endorsement. Corrections: cmmc20x@deepfathom.ai.