RESEARCH PAPER

The Minimum Viable CMMC 20X Pilot

A pilot should compare methods on frozen cases, publish dangerous errors, and earn each expansion of permitted use.

STATUS	LAST REVIEWED	RESEARCH LANES	CLAIM REGISTER
Current	**August 13, 2026**	**Evaluation · Mechanism · Current intervention**	**C-07 · C-08 · C-11**

Evaluation design for discussion. It proposes a no-reliance comparison and carries no assessment, contracting, certification, or program effect.

A credible pilot begins with a decision, rather than a demonstration.

The decision might be whether software-assisted review may help a qualified reviewer identify missing or conflicting support in a defined class of Level 2 evidence. It cannot be whether an AI system may certify a contractor. CMMC assigns findings, affirmations, certification, acquisition, and enforcement to authorized people and organizations. A pilot should preserve those authorities while testing a narrower method.

This distinction determines the whole study. A polished demonstration can prove that software runs. It cannot establish accuracy, safety, comparative burden, or acceptable use.

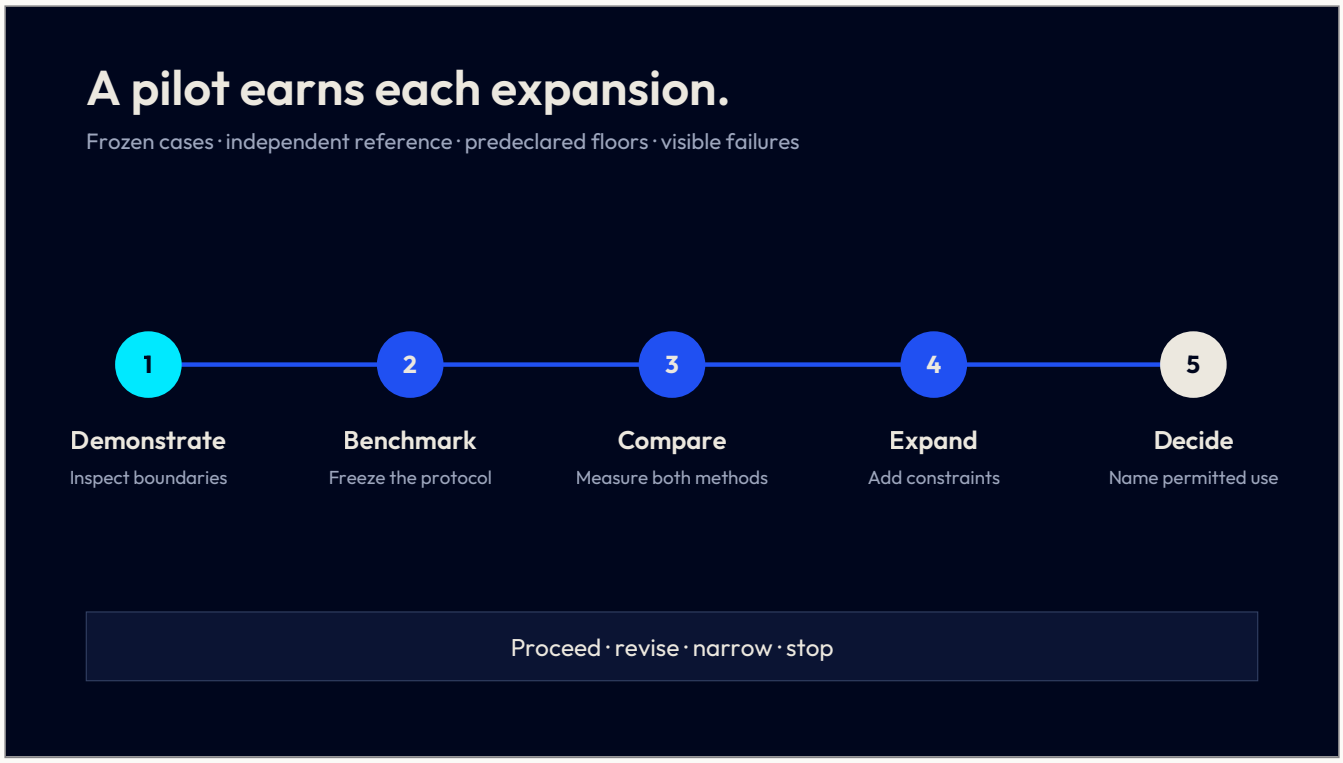
The smallest useful question

The minimum viable CMMC 20X pilot asks:

> On the same frozen cases, can software-assisted review surface unsupported, stale, conflicting, or out-of-scope claims sooner while maintaining predeclared quality and safety floors?

That question has five virtues. It names the task, the comparison, the failure classes, the expected benefit, and the condition that prevents speed from winning by concealing errors.

The first study should use synthetic cases. Synthetic does not mean easy. Cases should include technical exports, policies, interviews, provider dependencies, sparse telemetry, material changes, and deliberately misleading artifacts. Their advantage is control: every participant can receive the same record, and the reference panel can know which conflicts were planted.



The evaluation earns each expansion. A useful final decision may be proceed, revise, narrow, or stop.

Freeze before measuring

Changing a case or scoring rule after reviewers begin makes the comparison uninterpretable. The sponsor should freeze the protocol before scoring:

FREEZE	REQUIRED RECORD	REASON
Decision	Permitted conclusion and excluded conclusions	Prevents a benchmark from becoming an authorization
Cases	Version, digest, expected scope, planted conditions	Gives every method the same work
Reference	Independent findings with cited support	Provides a comparison target
Measures	Definitions, units, and aggregation rules	Prevents selective reporting
Floors	Minimum results for each relevant case group	Stops an average from hiding a dangerous subgroup
Stop rules	Data, integrity, safety, and protocol failures	Makes suspension automatic when a critical condition appears

Qualified reviewers should establish reference findings independently of Deep Fathom output. Reasonable disagreement should remain visible. For a judgment-dependent objective, the record may support more than one defensible finding. The benchmark should capture the source of that disagreement instead of forcing false unanimity.

Compare work, quality, and safety together

Elapsed time alone is a poor pilot result. A method can appear fast by skipping hard cases. Agreement alone is also weak when all reviewers inherit the same unsupported premise.

The pilot therefore needs three measurement families:

01 **Burden:** reviewer labor, supplier labor, duplicated handling, elapsed time, and rework.

02 **Quality:** coverage, traceability, supported findings, missed conflicts, reviewer agreement, and reproducibility.

03 **Safety:** false negatives, unsupported conclusions, unsafe reliance, resistance to misleading inputs, correction behavior, and results by case type.

Report distributions and case-level failures. A single average obscures whether a method performs well on identity exports and poorly on inherited service claims. Those cases create different risks and require different evidence.

Ordinary and assisted reviewers should work from the same cases. Instrument each action: what the reviewer opened, what a rule flagged, what an AI component suggested, what the person accepted or rejected, and why. This makes the method inspectable without treating activity logs as proof of sound judgment.

Dangerous errors deserve first-class status

A dangerous error is one that could create unjustified reliance: missing an access path that bypasses MFA, accepting evidence from the wrong system boundary, overlooking an expired provider responsibility, or converting an ambiguous source into a confident finding.

The protocol should declare which errors stop a run. It should also state which failures require narrowing a proposed use. A tool that reliably checks hashes may still be unsuitable for interpreting interview evidence. A component that maps artifacts may still require a person to decide whether the mapped evidence is sufficient.

Publishing only accuracy and time would give readers an incomplete result. Publish errors, corrections, disagreements, and stopped runs alongside favorable results.

Expansion must be earned

The pilot should progress through explicit gates:

```
demonstrate → benchmark → compare → expand under control → decide
```

The first gate shows the working system and its authority boundaries. The second establishes cases and reference findings. The third measures the proposed method against ordinary review. The fourth adds representative constraints only after the initial floors hold. The fifth produces a deployment decision that names permitted uses, required controls, excluded uses, and stop conditions.

There is no universal “pilot passed.” Results might support deterministic validation while rejecting AI-generated findings. They might support reviewer prioritization for technical evidence while requiring more work on provider inheritance. A useful decision can be proceed, revise, narrow, or stop.

What the pilot leaves behind

Even a negative result should produce public value:

- versioned benchmark cases another qualified team can rerun;
- reference findings with cited evidence and recorded disagreements;
- definitions for burden, quality, and safety;
- a catalog of failure modes and corrections;
- requirements for provenance, access, retention, and human authority;
- a clear account of the uses the evidence does and does not support.

That record is more valuable than a vendor score. It turns a product evaluation into reusable program knowledge.

CMMC 20X proposes that maintained, source-linked evidence and carefully evaluated software assistance may reduce repeated work and focus qualified reviewers. The minimum viable pilot is designed to find out where that proposition holds, where it fails, and what must remain under human control.

Inspect the full evaluation design, download the pilot charter, or follow the synthetic MFA case.

TRACE THE CLAIM

Sources and revision history.

PRIMARY AND GOVERNING SOURCES

01 **CMMC Assessment Guide — Level 2, Version 2.13**
Department of War Chief Information Officer
↗

02 **Artificial Intelligence Risk Management Framework**
National Institute of Standards and Technology
↗

03 **CMMC 20X Government Pilot Evaluation Charter v0.1**
Deep Fathom
↗

CORRECTIONS AND MATERIAL REVISIONS

No material corrections recorded.

See an error or a source we missed? Send a correction. Material changes are recorded here rather than silently overwritten.