INDEPENDENT EVALUATION

SOW/PWS Builder Testing Record

Federal acquisition skill validated across fourteen runs in two testing waves on two Claude models

Skill
sow-pws-builder (post-Wave 2 patches; Wave 3 validation pending)
github.com/1102tools/federal-contracting-skills

Date
April 2026
1102tools.com

Part 1: For Federal Acquisition Users

The bottom line

Two waves of independent testing in April 2026 (14 end-to-end runs, 168 binary assertions graded) show the SOW/PWS Builder reliably produces FAR-compliant Statements of Work and Performance Work Statements across six contract-type and workflow combinations on both Claude Opus 4.7 and Claude Sonnet 4.6. Five assertions failed across the two waves. Four of the five point to specific places where the skill's output needs a manual pass before solicitation release. None produced a document that was unsalvageable.

Contract types tested and how reliably they work

Contract type and workflow	Models	Result
T&M Statement of Work, non-commercial	Opus, Sonnet	Reliable
FFP PWS (commercial, full build)	Opus, Sonnet	Reliable
FFP PWS (commercial, SOO conversion)	Opus, Sonnet	Reliable
FFP PWS scope reduction (Workflow C)	Opus, Sonnet	Reliable after Wave 1 patch
CPFF R&D PWS, non-commercial	Opus, Sonnet	Reliable; verify CPFF form (see checklist)
Labor-Hour PWS, no materials, non-commercial	Opus, Sonnet	Reliable; verify DD Form 254 (see checklist)

Manual-verification checklist

The Wave 2 testing surfaced four specific gaps in the skill's output. **All four were patched after Wave 2.** The patches have not yet been regression-tested (Wave 3 will do that), so until the next wave confirms the fixes hold, scan every output for these items.

1. Key Personnel substitution clause. Historical gap: the skill emitted "FAR 52.237-2" as the Key Personnel clause on 5 of 6 Wave 2 runs. FAR 52.237-2 is actually "Protection of Government Buildings, Equipment, and Vegetation" — the wrong clause. The Wave 2 patch removed that default and added agency-aware guidance: NFS 1852.237-72 for NASA, HSAR 3052.237-72 for DHS, HHSAR 352.237-75 for HHS, DFARS where applicable for DoD, and a generic fallback that avoids FAR 52.237-2 entirely. Verify the citation still looks right for your agency.

2. CPFF form selection. Historical gap: on both Wave 2 CPFF runs, the skill cited FAR 16.306 without selecting completion form or term form. The Wave 2 patch added a Phase 2 rule requiring the skill to explicitly name completion form (FAR 16.306(d)(1)) or term form (FAR 16.306(d)(2)) wherever the contract framework first appears. Verify your CPFF PWS says which form it is. The difference matters: completion form requires delivery of a specified end product before the full fixed fee is earned; term form obligates a stated level of effort over a stated period. Term form is usually the right default for R&D where technical outcomes are uncertain.

3. Classified security block. Historical gap: on both Wave 2 Labor-Hour TS/SCI runs, DD Form 254 and Security Classification Guide references were missing. The Wave 2 patch added required DD 254 and SCG references in Section 9 whenever any clearance at Confidential or higher is called out. Verify the security block for any classified requirement.

4. Section ordering. Historical gap: workers occasionally swapped Section 11 (Reporting) and Section 12 (QASP) across runs. The Wave 2 patch added prescriptive language fixing Sections 11 through 14 in their correct order and explicitly disallowing merging or reordering. Low severity. A visual scan is enough.

Choosing between Opus 4.7 and Sonnet 4.6

Short answer: use Opus 4.7 as the default.

Opus 4.7 asks multi-choice clarification questions when the input has real ambiguity, waits for your explicit "proceed" before generating the document, and produces slightly more thorough clause citations. Two caveats:

- On claude.ai web chat, the Opus safety classifier sometimes blocks legitimate biodefense, vaccine, or pathogen-adjacent scenarios. If you hit a block ("Chat paused"), switch to Sonnet 4.6

or use the Claude API directly. This happened once in testing on an NIH mRNA vaccine R&D prompt.
- Opus consistently cites the wrong Key Personnel clause (see checklist item 1).

Sonnet 4.6 is faster and works well on richly specified prompts. Two caveats:

- Sonnet tends to skip the clarification-question phase even when defaults should have been reviewed. On the DHS Labor-Hour test, Sonnet self-approved six applied defaults without waiting for user confirmation. Read the Decision Summary carefully and interrupt if you see something that needs adjusting.
- On very large documents (roughly 30 KB and up), Sonnet on claude.ai web chat sometimes hits an output truncation limit mid-generation. A fresh chat retry usually completes.

For a critical document, run Opus first. If Opus is blocked or unavailable, switch to Sonnet.

What the skill does not do

- **It does not produce Independent Government Cost Estimates.** It emits a staffing handoff table in chat only, for the IGCE Builder skill to consume as a separate step.
- **It does not produce CLIN structures inside the PWS body.** CLINs go in Section B of the solicitation, and the skill correctly emits them as a chat-only handoff.
- **It does not handle classified data.** It produces unclassified acquisition documents that describe requirements for classified work. No actual classified information has been processed in testing.
- **It has not been tested on** Special Access Program or Special Access Required variants; commercial CPFF or commercial Labor-Hour (which require FAR 12.207(b) justifications); hybrid contract structures such as FFP plus T&M CLINs; multi-year ramping AQL structures where targets escalate across option years; or true document-to-document scope reduction using an actual prior .docx as input.

Environmental gotchas on claude.ai web chat

Gotcha	What happens	Workaround
Opus 4.7 biodefense classifier	Chat pauses mid-intake on vaccine, pathogen, or similar scenarios	Switch to Sonnet 4.6 or use Claude API
Sonnet 4.6 output truncation	"Claude's response could not be fully generated" on very large PWS builds	Fresh chat retry; optionally add size-constraint language to the prompt
Sonnet 4.6 silent self-approval	Skipped "proceed" gate on one run; applied defaults without confirmation	Read the Decision Summary carefully; interrupt to correct if needed

Part 2: For Developers and Technical Reviewers

Testing methodology

Two testing waves, both in April 2026. Same methodology both times.

- Worker sessions ran end-to-end in fresh claude.ai web chat, the same environment the skill's end users run in.
- A separate Claude Code Opus 4.7 instance (1M context window, max effort mode, Claude Max 20x subscription) graded each run independently.
- The grader had no access to the worker's drafting conversation or self-assessment until after grading was complete. This separation distinguishes "observed behavior" from "claimed behavior."
- Each run was graded against a 14-point binary assertion matrix (8 general assertions plus 6 scenario-specific assertions).
- All assertions were committed before seeing any worker output, to prevent the grading standard from drifting to accommodate whatever the worker happened to produce (same discipline as pre-registering a study).

Grading methodology

Each assertion received a binary pass or fail plus a one-line note on anything suspicious.

General assertions (8, unchanged across both waves):

- G1 Acquisition Strategy Intake collected before drafting
- G2 Phase 1 decision tree executed with visible Decision Summary before docx generation
- G3 .docx produced
- G4 No FTE counts, SOC codes, or staffing tables in document body (FAR 37.102(d))
- G5 No staffing or IGCE appendix in document
- G6 Staffing handoff chat-only, not saved as a file
- G7 Section numbering sequential
- G8 Key Personnel by role and qualifications, not by headcount

Scenario-specific assertions (6 per scenario): tailored to the contract-type and workflow behaviors being exercised.

Wave 1 (initial, pre-publication testing)

Six canonical scenarios plus two AskUserQuestion stress tests.

Wave 1 scenarios

- **Scenario 1:** Non-commercial T&M SOW, Workflow A, AFRL test range instrumentation under FAR Part 15 and FAR 16.601. Exercised the T&M Labor Category Ceiling Hours exception under FAR 16.601(c)(2).
- **Scenario 2:** Commercial FFP PWS, Workflow B (SOO conversion), Treasury Digital Customer Experience Modernization. Exercised SOO parsing, gap identification, and performance-based outcome framing.
- **Scenario 3:** FFP PWS Workflow C scope reduction, cybersecurity incident response and digital forensics. Reduced from $62M per year to $40M per year. Exercised trade-off menu generation, user validation gate, section regeneration, updated staffing handoff, and Section 14 cut documentation.
- **AskUserQuestion stress test:** One-sentence lazy prompt ("Need a PWS for help desk services for my agency. Can you help me write one?") to test whether the skill drives structured multi-choice intake when a non-expert user provides minimal context.

Wave 1 results

Opus 4.7 worker results

Scenario	Workflow	Assertions	Result
Non-commercial T&M SOW	A	14 of 14 pass	PASS
SOO-to-PWS Conversion	B	14 of 14 pass	PASS
Scope Reduction	C	14 of 14 pass	PASS

Sonnet 4.6 worker results

Scenario	Workflow	Assertions	Result
Non-commercial T&M SOW	A	14 of 14 pass	PASS
SOO-to-PWS Conversion	B	14 of 14 pass	PASS
Scope Reduction	C	13 of 14 pass	FAIL (G4)

AskUserQuestion feature test

Model	Behavior observed
Opus 4.7	Defaulted to prose questions on first pass. Reached for the structured multi-choice tool only when explicitly prompted by the user.
Sonnet 4.6	Auto-triggered the structured multi-choice tool on the first message. Batched 12 questions across 4 groups of 3. Included "Not sure" and "Something else" escape hatches. Produced a 422-paragraph contract-file-ready PWS after the user selected option one for every question.

The one Wave 1 bug

Sonnet Scenario 3 G4 failure. The Workflow C scope-reduction documentation in Section 14 contained specific staffing counts ("3-4 FTE examiners," "2-3 platform engineers," "8 Tier 2 analysts") as prior-state descriptors. Document bodies remain subject to FAR 37.102(d) even when

describing historical or prior state. Opus passed the same scenario using capability-framed language ("night-shift onsite headcount eliminated," "no dedicated playbook engineer") without specific counts.

Root cause: the skill's FAR 37.102(d) enforcement block was focused on staffing tables and the Phase 3 handoff. It did not explicitly address Section 14 scope-reduction narrative. Workers interpreted "describe what was cut" as license to quote the prior staffing plan.

Fix: added explicit guidance in the Section 14 block requiring capability-and-coverage language, not staffing counts, for both current-state and historical-state descriptions. Included compliant and non-compliant examples. Prescribed the five-field documentation structure (Prior Scope, Revised Scope, Estimated Annual Savings, Rationale, Residual Risk).

Wave 2 (follow-up, post-publication testing)

Three contract-type and workflow combinations not exercised in Wave 1. Each run once on Opus 4.7 and once on Sonnet 4.6, for six runs total.

Wave 2 scenarios

- **Scenario 4:** Commercial FFP PWS, Workflow A, GSA PBS enterprise IT help desk for 25,000 users across 11 regions, $15M per year, 24x7x365, NIST 800-171 and CMMC Level 2 required.
- **Scenario 5:** Non-commercial CPFF R&D PWS, Workflow A, NASA Glenn Research Center cryogenic fluid management research for in-space propellant transfer and long-duration storage, $45M with 8% fixed fee, 5-year term, Principal Investigator model, NASA FAR Supplement applies.
- **Scenario 6:** Non-commercial Labor-Hour PWS (no materials), Workflow A, DHS CISA cyber threat analyst advisory support to the Cyber Threat Intelligence Integration Office, $8M NTE over 3 years, TS/SCI for all positions, 5 labor categories.

Scenario 5 was originally planned as NIH/NIAID mRNA vaccine platform R&D. The Opus 4.7 safety classifier on claude.ai web chat paused the worker mid-intake. The scenario was swapped to NASA Glenn CFM (same contract mechanics, no biodefense trigger words) to preserve apples-to-apples grading across models.

Wave 2 results

Run	Model	Scenario	Result	Failed assertion
1	Opus 4.7	Commercial FFP (GSA PBS)	14 of 14 PASS	—
2	Sonnet 4.6	Commercial FFP (GSA PBS)	14 of 14 PASS	—
3	Opus 4.7	CPFF R&D (NASA Glenn)	13 of 14 PASS	S5
4	Sonnet 4.6	CPFF R&D (NASA Glenn)	13 of 14 PASS	S5
5	Opus 4.7	Labor-Hour (DHS CISA)	13 of 14 PASS	S6
6	Sonnet 4.6	Labor-Hour (DHS CISA)	13 of 14 PASS	S6

The two Wave 2 failure modes

Failure mode A: S5 on CPFF, both models. Neither worker explicitly committed the document to completion form (FAR 16.306(d)(1)) or term form (FAR 16.306(d)(2)). Both documents cited FAR 16.306 multiple times but never made the form selection explicit. The distinction governs how fixed fee is earned: completion form ties earning to delivery of a specified end product; term form ties earning to level of effort over a stated time period. For R&D where technical success is uncertain, term form is usually more appropriate.

Root cause: skill-template gap. The Phase 2 CPFF block does not currently require or emit an explicit form selection.

Candidate patch: add a Phase 2 rule requiring the document to name completion form or term form wherever the contract framework first appears.

Failure mode B: S6 on Labor-Hour, both models. Neither worker referenced DD Form 254 in the security section of a TS/SCI Labor-Hour PWS. Both covered clearance levels, SCIF requirements per ICD 705, and derivative classification, but DD Form 254 and Security Classification Guide references were missing.

Root cause: skill-template gap. The Phase 2 security block does not currently emit DD Form 254 or SCG references for classified requirements.

Candidate patch: when clearance above Confidential is called out, the Phase 2 security block must reference DD Form 254 and the applicable SCG.

Behavioral observations from Wave 2 (not assertion failures)

1. **FAR 52.237-2 misreference for Key Personnel substitution.** Appeared on 5 of 6 Wave 2 runs. Sonnet corrected it to NFS 1852.237-72 on the NASA CPFF scenario but reverted to the template default on the other runs. Skill-template bug, not a worker variance.
2. **Sonnet skips AskUserQuestion intake on medium-ambiguity prompts.** Sonnet triggered AskUserQuestion on 0 of 3 Wave 2 scenarios. Opus triggered on 2 of 3. This inverts the Wave 1 finding, where Sonnet was the auto-trigger model and Opus needed explicit prompting. The Wave 1 patch fixed Opus; Sonnet has now drifted in the other direction.
3. **Sonnet self-approved Phase 1 Decision Summary on Scenario 6.** Said "Proceeding to document assembly now" immediately after presenting the summary, without waiting for user "proceed." This denies the user a chance to override applied defaults before generation. The G2 assertion (visible Decision Summary before docx) passed on visibility, but the user-gate is a separate concern.
4. **Section ordering drift.** QASP and Transition sections swap positions (Section 11 versus Section 12) across and within models. Terminology is always correct; only position drifts. Low severity.
5. **Opus 4.7 biodefense classifier block.** Environmental observation on claude.ai web chat, not a skill defect. The classifier paused the worker mid-intake on an NIH/NIAID mRNA vaccine R&D scenario. Federal acquisition users working on biodefense, vaccine, or pathogen-related contracts should expect potential classifier blocks on Opus web chat and plan to use Sonnet or the Claude API.
6. **Sonnet 4.6 output truncation on large documents.** Environmental observation. First attempt on Scenario 4 Commercial FFP PWS hit "Claude's response could not be fully generated" mid-.docx-generation. Fresh chat retry completed cleanly.
7. **Sonnet domain reasoning on CPFF was notably sharp.** Sonnet on Scenario 5 caught that CPFF fee cannot be administratively reduced under FAR 16.306 and flagged the skill's QASP fee-deduction language as legally problematic, recommending either repositioning as a cure-notice tool or conversion to CPAF/CPIF. That is substantive acquisition-law reasoning, not template output. Opus did not surface this on its parallel CPFF run.

Cumulative results across both waves

	Wave 1	Wave 2	Total
Worker runs	8 (6 canonical + 2 AskUserQuestion)	6	14
Scenario assertions graded	84	84	168
Passed	83	80	163
Failed	1 (fixed in patch)	4 (cross-model, patch candidates)	5

Skill patches: shipped and candidate

Shipped after Wave 1 (9 patches)

Patch	Section affected	Trigger
Unconditional handoff rule moved to top of Phase 3 with emphatic language	Phase 3 opening	3 of 7 runs skipped handoff emission
SOW "Inspection and Acceptance" versus PWS "QASP" label differentiation	Phase 2 Section 12 block	Scenario 1 both models
Table of Contents instruction for documents exceeding 8 sections	Phase 2 opening	All runs produced 300-700 paragraph docs without TOC
SOO-implied objectives rule	Phase 0	Scenario 2 both models added unstated objectives legitimately
Phase 2 Invocation Gate requiring Phase 1 Decision Summary before docx	Phase 2 opening	Cross-model Phase 1 compression plus network-blip resilience
Section 14 assumption format template (4-column)	Phase 2 Section 14 block	Workers independently invented divergent structures
Workflow C Section 14 compliance rule (no FTE counts in cut descriptions)	Phase 2 Section 14 block	The Wave 1 G4 failure
AskUserQuestion tool usage instruction	Phase 1 opening	Opus defaulted to prose without explicit instruction
Anti-redundancy rule	Phase 1 opening	Both models re-asked explicit answers

Skill version lines: 361 before Wave 1 patches, 380 after. Ceiling: 1,000.

Shipped after Wave 2 (5 patches)

Patch	Section affected	Trigger
Agency-aware Key Personnel substitution clause selection (removed the wrong FAR 52.237-2 default; added agency-specific guidance for NASA, DHS, HHS, DoD, and a generic fallback)	Phase 2 Key Personnel block	FAR 52.237-2 emitted on 5 of 6 Wave 2 runs
Explicit CPFF form commitment required (skill must name completion form per FAR 16.306(d)(1) or term form per FAR 16.306(d)(2) wherever contract framework first appears)	Phase 2 Section 5 block	S5 failure on both Wave 2 CPFF runs
DD Form 254 and SCG reference added to Section 9 for any classified requirement at Confidential or higher	Phase 2 Section 9 block	S6 failure on both Wave 2 LH runs
Strengthened Phase 1 proceed gate; explicit "DO NOT self-approve" rule with requirement that the response must END after presenting the Decision Summary	Phase 2 Invocation Gate	Sonnet self-approved on Scenario 6
Fixed section ordering made explicit and prescriptive (Section 11 Reporting, Section 12 QASP, Section 13 Transition, Section 14 Constraints — do not merge, combine, swap, or rename)	Phase 2 Section Structure header	QASP and Transition position drift across runs

Skill version lines: 426 before Wave 2 patches, 450 after. Ceiling remains 1,000.

These patches are shipped in the current skill but have not yet been validated against a fresh test wave. Wave 3 will regression-test the same six scenarios against the patched skill to confirm the Wave 2 failure modes no longer reproduce.

What was not tested

- **Special Access Program and Special Access Required variants.** Classified coverage was limited to collateral Secret, Top Secret, and TS/SCI contexts in notional scenarios.
- **Commercial CPFF or commercial Labor-Hour.** Both require FAR 12.207(b) written Determinations and Findings and are edge cases on top of edge cases.

- **Hybrid contract structures** (FFP plus T&M or LH CLINs; FFP with cost-reimbursable travel CLINs was exercised only incidentally in the Wave 2 handoff tables).
- **True document-to-document scope reduction** using an actual prior .docx as Workflow C input. The Wave 1 test used a bulleted summary of the prior PWS rather than an actual document.
- **Multi-year ramping AQL structures** where performance targets escalate across option years. Exercised informally in Wave 1, not systematically.
- **Hardware-intensive Test and Evaluation domains** beyond the single Wave 1 Scenario 1 (AFRL test range) and Wave 2 Scenario 5 (NASA CFM).
- **Grants, cooperative agreements, and Other Transaction agreements** are out of scope for this skill. Separate skills cover those.

Users working in these contexts should expect to validate outputs more carefully and may encounter edge cases that the test waves did not surface.

Note on classified scenario testing

Wave 1 Scenario 1 used a notional classified-context prompt (AFRL test range, Secret and Top Secret collateral clearances). The test verified only that the skill produces proper classified-contract boilerplate: clearance-by-position requirements, Security Classification Guide references, OPSEC language, facility clearance statements, and DD Form 254 references at the prompt level. All scenario inputs were fictional. No actual classified information was processed, transmitted, or stored at any point during testing. The skill itself does not and cannot handle classified data; it produces unclassified acquisition documents that describe requirements for classified work, which is standard practice for any SOW or PWS.

Wave 2 Scenario 6 (DHS CISA TS/SCI) extended classified-context coverage to a TS/SCI Labor-Hour PWS using a fictional CISA scope. Same constraint: no real classified information was ever handled.

Coverage was limited to collateral Secret, Top Secret, and TS/SCI contexts. Special Access Program and Special Access Required variants were not exercised.

Independent grading

Every assertion was graded by a separate Claude instance reading only the worker's final output and the chat transcript. The grader had no access to the worker's internal reasoning and did not see the worker's self-assessment until after grading was complete. This separation is the load-bearing credi-

bility claim of this testing program: it distinguishes "the worker claimed to do X" from "the output demonstrates X."

> **Testing Methodology**
>
> Evaluators: James Jenrette (1102tools) and Claude Code Opus 4.7 (1M context window, max effort mode, Claude Max 20x subscription).
>
> Worker models tested: Claude Opus 4.7 and Claude Sonnet 4.6 on claude.ai web chat, the same environment the skill's end users run in.
>
> Wave 1: 8 runs, 84 assertions, 83 passes, 1 confirmed failure (fixed in patch).
>
> Wave 2: 6 runs, 84 assertions, 80 passes, 4 confirmed failures (cross-model, patch candidates identified).
>
> Cumulative: 14 runs, 168 assertions, 163 passes.
>
> Date: April 2026 (both waves).
>
> Skill: sow-pws-builder. Source: github.com/1102tools/federal-contracting-skills. License: MIT.