

---

## TESTING RECORD

# OT Project Description Builder: Testing Record

---

*One wave of live testing across five OT scenarios in April 2026: NDC prototype, SOW conversion, BAA consortium with competition commitment, and research OT. 15 universal patches shipped, zero regressions on post-patch runs.*

---

### Skill

ot-project-description-builder  
[github.com/1102tools/federal-contracting-skills](https://github.com/1102tools/federal-contracting-skills)

### Date

April 2026  
[1102tools.com](https://1102tools.com)

## The bottom line

One testing wave in April 2026 across six end-to-end scenarios on Claude Opus 4.7 shipped 15 patches to the OT Project Description Builder. All 15 patches validated on regression across three post-patch runs that exercised untested territory (software prototype, consortium, path D competition commitment, research OT, traditional prime, and Workflow C scope reduction). No regressions, no new universal gaps. All three workflows (A full build, B document conversion, C scope reduction) are now covered.

The skill reliably produces 10 USC 4021 and 4022 compliant project descriptions across prototype, research, and production-follow-on paths, with milestone-based payment structures, TRL-mapped phase progression, data rights handling, cost-sharing path selection, and a separate chat-only milestone handoff table for the OT Cost Analysis.

## Scenarios tested

#	Scenario	Authority	Performer path	Workflow
1	Ultra-lazy drone prototype	10 USC 4021	NDC	A (full build)
2	Tethered surveillance drone, richly specified	10 USC 4021	NDC, 4022(d)(1)(A)	A
3	FAR-flavored SOW conversion with CLIN/cost contamination	10 USC 4021	Traditional, 4022(d)(1)(C)	B (conversion)
4	NSTXL-brokered BAA white paper, software prototype	10 USC 4021 + 4022(f)	Traditional with competition commitment, 4022(d)(1)(D)	B
5	Cold-weather battery chemistry research	10 USC 4021 (research)	Traditional, no cost-share required	A
6	Autonomous resupply prototype, \$18M priced, \$11M approved, descope to fit budget	10 USC 4021	Traditional, 4022(d)(1)(C)	C (scope reduction)

All six runs produced contract-file-ready .docx outputs, emitted the Milestone Handoff Table as chat output only, and stopped at the Phase 2 Invocation Gate for user "proceed" before document generation.

## What the skill got right on every run

- Four-question Acquisition Context Intake up front (OT type, performer type, TRL entry/exit, consortium/direct)
- Phase 2 Invocation Gate held. No self-approval observed on any run after the gate patch landed.
- Milestone Handoff Table presented as chat-only markdown, never embedded in the .docx or saved as a separate file
- Section 12 Cost-Sharing conditionally omitted for NDC, SB, competition-commitment paths, and for 10 USC 4021 research authority (where 4022(d) is statutorily inapplicable)
- No FAR clause references leaked into the body. The one permitted exception (DFARS 252.227-7013/7014 cited as data rights taxonomy framework under OT tailoring) handled correctly.
- No cost figures, CLINs, FTE counts, or SOC codes in any document body
- Traditional SOW-to-OT conversion correctly stripped CLIN structure, FAR 52 clauses, QASP/AQL language, and dollar amounts while preserving scope intent

## Patches shipped in this wave

Patches were applied in two sub-groups. Group 1 was cross-shipped from the SOW/PWS Builder Wave 2 patch set. Group 2 was derived from three live OT runs and validated against two subsequent runs covering untested territory.

**Group 1: Cross-shipped from SOW/PWS Wave 2 (5 patches)**

Patch	Trigger
Phase 2 Invocation Gate with "DO NOT self-approve" language	Universal pattern in document-generating skills
Anti-redundancy rule (do not re-ask questions the user's prompt already answers)	Universal pattern
AskUserQuestion guidance for Phase 1 intake batching	Matches claude.ai web chat behavior
UNCONDITIONAL RULE requiring Milestone Handoff Table emission	Mirrors SOW/PWS unconditional handoff pattern
Section ordering prescriptive with explicit renumbering rule for omitted conditional sections	Prevents drift across runs

## Group 2: OT-specific universal gaps surfaced in Tests 1-3 (10 patches)

Patch	Section affected	Trigger
Corrected NDC definition to CAS full-coverage test under 10 USC 3014, removed the incorrect "\$500K+ in prior year" threshold	Intake Q2	Test 2 flagged the factual error; verified against statute
Conditional docx generator path (use /mnt/skills/public/docx/SKILL.md on claude.ai sandbox, fall back to python-docx on Claude Code and local installs)	Phase 2 opening	Test 2 flagged the sandbox-only path; all three tests had workarounds
Split canned handoff closing message into two variants (cost-share and no-cost-share)	Phase 3 handoff	The single version incorrectly promised "apply cost sharing" on NDC, SB, and competition paths
Required Document Review Checklist emission in chat (with per-item pass or attention status) before the Milestone Handoff Table	Phase 3	All three initial runs ran the checklist mentally and moved on
Go/no-go gate carry-through rule (phase-boundary milestones must carry Block 2 Q7 criteria into both Section 4 and Section 5)	Phase 1 milestone derivation	Test 1 flagged; criteria collected but not surfaced in the document
Defined "Est. Duration" column semantics (calendar months from prior milestone)	Phase 3 handoff table	Ambiguous durations produce 3x+ cost-analysis labor loading deltas
Cost data stripping rule for Workflow B (strip source dollar figures from body, pass to handoff as informational only)	Phase 0 Document Intake	Test 3 needed this; skill was silent

Patch	Section affected	Trigger
Cost-share arithmetic disambiguation for Workflow B + 10 USC 4022(d)(1)(C)	Phase 0 Document Intake	Whether a stated total is government share or total agreement value is ambiguous in most SOWs and produces materially different government obligations
Performance placeholder closeout mechanism for [TBD] thresholds in Section 3 Objectives, with default bilateral-modification disposition at Phase 1 CDR	Phase 2 Section 14	Multiple runs invented disposition mechanisms; standardized here
Body cross-references must use section titles rather than section numbers	Phase 2 Section Structure	Prevents broken refs when conditional sections are omitted and subsequent sections renumber

## Regression validation (Tests 4, 5, and 6)

Tests 4, 5, and 6 exercised untested territory after all 15 patches landed. They were designed specifically to stress parts of the skill that Tests 1-3 didn't touch.

- **Test 4 (BAA conversion, consortium, path D, software, late TRL 5-7):** Every Group 2 patch that could fire did fire correctly. The conditional handoff message used the no-cost-share variant. The Document Review Checklist emitted as 11 items in chat. Cross-reference patch held (no numbered refs in body). Source-doc cost stripping wrote "None. White paper contained no cost figures" to the handoff, demonstrating the rule holds even when there is nothing to strip.
- **Test 5 (Research OT, traditional prime, 30 months):** Skill correctly identified that 10 USC 4022(d) cost-sharing paths do not apply to 10 USC 4021 research authority and omitted Section 12. Cost-share arithmetic question did not fire because the statutory framework makes it inapplicable (the correct behavior, not a miss). Go/no-go criteria carried through as placeholder thresholds with stated disposition.
- **Test 6 (Workflow C scope reduction, traditional prime, path C cost-share):** Skill correctly routed to Workflow C. The cost-share arithmetic disambiguation patch, originally written for Workflow B, generalized cleanly to Workflow C and fired as a blocking question with three specific arithmetic interpretations (total value, government share, or pre-share performer

quote) and correct math for each. The source-doc cost stripping rule also generalized: original \$18M baseline and \$11M approved funding carried into the Milestone Handoff Table as "informational only, do not anchor should-cost estimate" rather than leaking into the regenerated document body. Document Review Checklist, Phase 2 Invocation Gate, conditional handoff message (cost-share variant), and [TBD] placeholder closeout all fired correctly. Model also showed useful domain judgment beyond skill rules (flagged NTC as TRL 7 operational and redirected to YPG for TRL 6 relevant environment, presented descope trade-offs as ranked packages A/B/C with risk levels). Those judgment moves were not patch-worthy.

No regressions observed across any of the three regression tests. No new universal gaps surfaced.

**Cross-workflow patch generalization.** Two Group 2 patches originally written for Workflow B (cost-figure stripping, cost-share arithmetic disambiguation) generalized to Workflow C without modification. This is the pattern we want: universal rules that apply wherever source cost data and path (C) math appear, not workflow-specific hedges.

## What was not tested

- **Production follow-on OT as primary workflow.** Tests 2 and 4 included 10 USC 4022(f) provisions as an option, but neither run structured the agreement as a production follow-on.
- **Consortium-specific templates.** Test 4 used NSTXL as the broker; DIU, AFWERX, NavalX, SOSSEC, and MTEC consortium variations have not been exercised.
- **Small business performer path (10 USC 4022(d)(1)(B)).** Tests covered NDC (path A), traditional with competition commitment (path D), traditional with cost-share (path C), and research (no 4022(d) applicability). Small-business significant participation was not tested.
- **Multi-performer consortia.** All tests used single performers or single primes with no subs.
- **Deep classified prototype contexts.** No SAP, SAR, or ICD 705 variants were exercised.
- **Hardware prototype beyond drone and ground vehicle.** Tests 1, 2, and 3 all touched airborne or ground-mobility prototypes. Directed energy, hypersonics, biotech, cyber tooling, and AI/ML prototypes beyond the Test 4 software case have not been exercised.

Users in these contexts should expect to validate outputs more carefully and may encounter edge cases this wave did not surface.

## Testing methodology

Six live runs on Claude Opus 4.7 via claude.ai web chat and Claude Code, the same environments the skill's end users run in.

Evaluator grading was done by a separate Claude Opus 4.7 instance with access only to the skill text and the worker's final output transcript. Grading focused on three questions per run:

1. Did each patch under test fire correctly?
2. Did the output surface any new structural gap not covered by existing patches?
3. If a gap was observed, is it a universal pattern across the skill's full usage profile, or a one-off of model judgment on a specific input?

Only gaps classified as universal structural patterns produced new patches. Drone-specific feedback (unit count defaults, airworthiness prompting), DoD-specific feedback (CMMC, ITAR), and over-engineering suggestions (sentinel markers, trigger-phrase hardening) were deliberately skipped to prevent bloat.

Skill line count: 408 before the wave, 440 after all 15 patches (+32 lines net on a skill that runs hundreds to thousands of times).

## Patches: shipped in this wave

Skill version lines: 408 before patches, 440 after. Ceiling remains 1,000.

All 15 patches are live in the current SKILL.md.



**Testing Methodology**

Evaluator: James Jenrette (1102tools) and Claude Code Opus 4.7 (1M context window, max effort mode, Claude Max 20x subscription).

Worker model tested: Claude Opus 4.7 on claude.ai web chat and Claude Code, the same environments the skill's end users run in.

Wave: 6 runs, 15 patches shipped, 3 post-patch regression runs with zero new gaps surfaced. All three workflows (A, B, C) covered.

Date: April 2026.

Skill: ot-project-description-builder. Source: [github.com/1102tools/federal-contracting-skills](https://github.com/1102tools/federal-contracting-skills). License: MIT.