

---

## TESTING RECORD

# IGCE Builder FFP Testing Record

---

*Nine waves of testing across April 2026 validated the Firm-Fixed-Price IGCE skill through 18 end-to-end runs plus cross-skill universal patches, including Wave 9 DCAA/FPRA override rule and Workflow B gate hardening ported from LH/T&M Wave 5.*

---

### Skill

igce-builder-ffp  
[github.com/1102tools/federal-contracting-skills](https://github.com/1102tools/federal-contracting-skills)

### Date

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

# Part 1: For Federal Acquisition Users

---

## The bottom line

Four waves of independent testing across April 2026 (18 end-to-end runs on Claude Opus 4.7 and Claude Sonnet 4.6, 210 binary assertions graded plus three MCP-era qualitative rounds) show the IGCE Builder FFP skill reliably produces auditable Firm-Fixed-Price cost estimates across seven distinct federal acquisition scenarios.

- **Wave 1** (claude.ai, pre-patch): 55/56 = 98% on Opus with Sonnet parity. Surfaced 17 cross-run quality issues.
- **Wave 2** (claude.ai, post-patch): 56/56 = 100% on Opus after 17 substrate patches.
- **Wave 3** (claude.ai, post-Round 5/6): 42/42 = 100% on Opus after burden-band recalibration and 6 additional patches.
- **Wave 4** (Claude Code CLI, post-MCP migration): three rounds on Opus 4.7 against the reduced skill orchestrating 3 MCP servers directly (bls-oews, gsa-calc, gsa-perdiem). 3 workbooks produced, 18 additional skill issues identified and patched. Zero workbook build failures, zero silent-wrong-answer bugs.
- **Wave 5** (Claude Code CLI, ai-boundaries + untested workflow paths): three rounds on Opus 4.7 covering Workflow A+ SOW decomposition (S4), multi-location explicit headcount (S5), and Workflow B rate validation (S6). Surfaced a Tier-1 ai-boundaries violation (Workflow B originating fair-and-reasonable determinations), the DoD installation → GSA per diem crosswalk gap, 22 additional skill issues. All patched.

## Scenarios tested and how reliably they work

Scenario	Models	Result
Standard FFP-by-period IT services build (DC dev team, base + 2 OYs)	Opus, Sonnet	Reliable both waves
24x7 shift coverage (Cleveland SOC analysts, base + 2 OYs)	Opus, Sonnet	Wave 1: Opus burned 15 min brute-forcing Cleveland MSA due to 2024 OMB renumbering from 17460 to 17410, shift math reconstructed from first principles. Wave 2: reliable after patches
Physical engineering multi-LCAT (Oak Ridge DOE environment)	Opus, Sonnet	Wave 1: SOC mapping defaulted to IT codes for Mechanical/Electrical Engineers, PM mapped to 11-3021. Wave 2: reliable after 17-2xxx block added and PM SOC made context-dependent
FFP-by-deliverable multi-milestone study (DISA feasibility analysis)	Opus, Sonnet	Wave 1: hour allocation across deliverables inconsistent (4/4 runs chose different methods), CALC+ rate validation hit silent-wrong-answer bug (q= vs keyword=) in 4/4 Opus runs. Wave 2: reliable after explicit allocation methods + CALC+ query signature inlined

## Manual-verification checklist

Scan every output for these before using in a contract file:

**1. CALC+ rate validation was actually validated, not faked.** The Wave 1 silent-wrong-answer bug came from the CALC+ API accepting q= and returning the full 265K-record corpus without error. If the rate validation sheet cites a median derived from "hundreds of matches" or variance exceeds 60% across the sample, the worker hit the bug. Patched skill inlines keyword= and /v3/api/ceilingrates/ endpoint; workers cannot accidentally route through the broken signature.

**2. Fully burdened rate must tie to BLS + wrap buildup, not just to CALC+ directly.** CALC+ reports awarded ceiling rates for completed MAS tasks. Using a CALC+ median as your direct FFP rate skips the buildup audit trail FAR 15.404-1(a) expects. Correct flow: BLS base → fringe

→ labor+fringe → overhead → subtotal → G&A → total cost → profit = FBR. Then compare FBR to CALC+ for reasonableness.

**3. 24x7 coverage needs 4.2 FTE per seat, not 1.** The skill's Step 0.5 computes 2,080 productive hours × 4.2 FTE to cover 8,760 annual hours (24 × 365) with leave/training/coverage slack. If a Cleveland SOC workbook shows 3 FTE for 24x7 coverage, the shift math is wrong. Double-seat (two analysts always on) is 8.4 FTE.

**4. Rate validation band should be 0-40% above CALC+ median for FFP.** Wave 2 patch calibrated: 0-15% is expected range. 15-40% is the FFP premium band (risk-adjusted fixed pricing justifies a markup over the ceiling-rate median). Above 40% needs explicit justification in the narrative. The pre-patch skill flagged anything over 10% which would have fired on nearly every legitimate FFP build.

**5. MSA renumbering silently returns empty data.** BLS returns the same "series does not exist" for a truly unpublished occupation and for a renumbered metro. If a metro query returns NO\_DATA across every SOC, check the OMB Bulletin 23-01 renumbering list before falling back to state. Cleveland (17460→17410), and possibly Dayton, shifted.

**6. Implied multiplier tells you whether the build makes sense.** Fully burdened rate divided by BLS base wage should land in the 2.2x-3.5x band for MID scenario. Below 2.2x means an unrealistic wrap assumption. Above 3.5x needs SCIF/OCONUS/niche justification. The HIGH scenario legitimately exceeds 3.5x and should not be flagged.

## Choosing between Opus 4.7 and Sonnet 4.6

Short answer: both work. Use either on the patched skill. Differences are small.

**Opus 4.7** handles multi-location builds and FFP-by-deliverable decomposition more reliably. In Wave 1, Opus caught the Cleveland MSA renumbering by brute-force scanning BLS; Sonnet fell back to state-level wages without flagging the metro issue. Opus is the preferred model for: SOW-driven builds where decomposition judgement matters, FFP-by-deliverable with per-LCAT hour matrices, any build touching a DOE or DoD specialty metro, any multi-LCAT build with 5+ labor categories. Opus also tool-uses more aggressively and sometimes hits the per-response tool-use cap on complex builds. If that happens, click continue.

**Sonnet 4.6** is faster on standard single-location FFP-by-period builds and produces cleaner workbooks in fewer tokens on the happy path. For a straightforward "3 FTE in DC, base + 2 OYs" job, Sonnet wins. Sonnet is less reliable at: metro-code validation when BLS returns empty, FFP-by-de-

liverable hour allocation decisions, and rate validation narrative text (tends to state rather than justify).

Wave 2 was run on Opus only for time reasons; Sonnet parity on the patched skill is inferred from Wave 1 where Sonnet matched Opus on 3 of 4 scenarios.

## What the skill does not do

- **It does not produce LH/T&M or cost-reimbursement estimates.** Use IGCE Builder LH/T&M or IGCE Builder CR respectively. The wrap rate buildup in FFP does not apply to those contract types.
- **It does not estimate subcontractor costs.** If the prime proposes 30% subcontract, you need separate vendor input or a second IGCE for the sub's scope.
- **It does not negotiate fee/profit.** It produces a cost buildup. Fee negotiation is a separate FAR 15.404-4 activity.
- **It does not handle OCONUS per diem.** GSA Per Diem covers CONUS only; use State Department rates for OCONUS assignments.
- **It does not price SCIF build-out, TEMPEST, or COMSEC equipment.** These require agency-specific quotes.
- **It has not been tested on:** CR-to-FFP conversion modeling (pricing legacy cost-plus work as FFP), FFP with award fee overlays (hybrid structures), ANSI/EIA-748 EVMS-compliant cost buildup formatting, international labor (BLS is US-only), or DCAA forward-pricing rate audits (the skill estimates cost, it does not audit vendor rate proposals against DCAA-disclosed rates).

## Environmental gotchas on claude.ai web chat

Gotcha	What happens	Workaround
Multi-LCAT build with 5+ categories + per-deliverable matrices	Opus hits per-response tool-use cap mid-build	Click continue; the skill resumes without repeating prior work
Complex workbook (9 sheets + rate validation dual-pool)	Python execution on claude.ai can time out	Ask model to build each sheet incrementally and present sheet-by-sheet
xlsx output doesn't appear in chat	File is in sandbox but not surfaced to UI	Patched skill's Step 9 explicitly calls <code>present_files()</code> after copying to <code>/mnt/user-data/outputs/</code>
CALC+ dual-pool query returns inconsistent record counts	Title-match and experience-match pools sometimes overlap	Use the patched Step 4 decision tree: title-match first, then experience-match as sanity layer if title match $N < 10$

# Part 2: For Developers and Technical Reviewers

## Testing methodology

### Scenarios

Four scenarios were selected before any testing began, chosen to exercise distinct capabilities across FFP pricing structures and federal agency contexts:

- **S1 — Standard FFP-by-period IT services:** 3 FTE developer team in DC, base year + 2 option years. Exercises SOC mapping for Software Developer (15-1252) + Senior Developer + Business Analyst, BLS DC metro (47900), full wrap buildup (35% / 85% / 10% / 10% MID), 2.5% escalation year-over-year, CALC+ rate validation at 520 SIN.

- **S2 — 24x7 shift coverage, specialty market:** Cleveland SOC analyst coverage 24x7x365, base + 2 OYs. Exercises shift coverage math (single-seat = 4.2 FTE), Information Security Analyst SOC (15-1212) BLS lookup, Cleveland MSA boundary (exposes 2024 OMB renumbering from 17460 → 17410), rate validation for an above-median SOC in a below-median metro.
- **S3 — Physical engineering multi-LCAT:** Oak Ridge TN DOE environment, 6-category staffing (Mechanical Engineer, Electrical Engineer, Chemical Engineer, Technical Writer, PM, Admin), base + 4 OYs. Exercises 17-2xxx SOC block (Wave 1 missed: skill defaulted Mechanical Engineer to 15-1211), PM SOC context-dependent (should be 11-9041 Engineering Manager, Wave 1 picked 11-3021 IT Manager), Oak Ridge MSA (28940), multi-LCAT wrap buildup, CALC+ dual-pool rate validation on senior engineering LCATs.
- **S4 — FFP-by-deliverable, multi-milestone study:** DISA 18-month feasibility study, 4 deliverables at 15/30/25/30 scope weights, SOW-driven build (Workflow A+). Exercises Step o requirements decomposition + validation gate, FFP-by-deliverable hour allocation across milestones (three valid methods: uniform split, per-LCAT matrix, staffing-profile), aging wages once to contract start with no mid-contract escalation, Summary sheet columns = CLINs.

Each scenario had a 14-point binary assertion matrix. Assertions were written before any worker output was seen and were not revised after the fact.

## Environment

- claude.ai web chat, fresh conversation per run
- Skills installed: `igce-builder-ffp` plus downstream `substrate` `bls-oews-api`, `gsa-calc-ceilingrates`, `gsa-perdiem-rates` (all merged, post-Round 2 patches for downstream skills)
- Models: Wave 1 ran each scenario on Opus 4.7 and Sonnet 4.6. Wave 2 ran Opus 4.7 only (time-constrained)
- Total: Wave 1 =  $4 \times 2 = 8$  runs. Wave 2 =  $4 \times 1 = 4$  runs. Aggregate = 12 runs / 168 assertions

## Grading

The grader (Claude Code session separate from any worker run) read only the worker's final response text and produced workbook. Workers were not coached during runs. Each assertion graded binary pass/fail. Suspicious details were noted even when assertions passed. Hour allocation ambiguity in S4 was graded as "worker picked one valid method and stayed internally consistent" rather than prescribing a specific method.

Wave 1 results (pre-patch)

Scenario	Sonnet 4.6	Opus 4.7
S1 DC dev team FFP-by-period	14/14	14/14
S2 Cleveland 24x7 SOC	13/14	14/14
S3 Oak Ridge DOE multi-LCAT	14/14	14/14
S4 DISA FFP-by-deliverable	14/14	13/14
Total	55/56 (98%)	55/56 (98%)

Wave 1 aggregate: 110/112 (98%).

Failures observed

**S2.X Sonnet — Cleveland 24x7 shift math:** Sonnet computed 3 FTE for single-seat 24x7 coverage. Correct is 4.2 FTE (8,760 annual coverage hours / 2,080 productive hours × availability factor for leave/training/turnover). The final workbook understaffed by 28% and the FFP total was commensurately low.

**S4.X Opus — CALC+ rate validation returned meaningless results:** Opus sent CALC+ queries using q= parameter. CALC+ accepted silently and returned the full 265K-record corpus. Rate validation narrative cited a "median of \$142.85 across 15,000+ matches" which was meaningless (population median, not occupation-specific). Workbook shipped with a broken validation sheet that looked fine on casual inspection.

The CALC+ bug was NOT unique to S4. Targeted re-inspection showed all 4 Opus Wave 1 runs hit this bug to varying degrees. S1, S2, S3 Opus runs still passed their rate-validation assertions because the grader checked "rate validation sheet exists with a median and a variance band" rather than "the median is arithmetically defensible." The patched assertion text for Wave 2 required the worker to cite the exact endpoint (/v3/api/ceilingrates/) and the exact parameter (keyword=) in methodology notes, forcing the query signature to be demonstrated rather than merely claimed.

Wave 1 findings: 17 cross-run issues patched

From Wave 1 worker self-assessments, grader notes, and cross-run observation:



1. **CALC+ query signature silently wrong in 4/4 Opus runs.** q= returns the full corpus; keyword= returns the filtered set. The CALC+ skill documentation showed q= in one example. The downstream skill was patched separately (see GSA CALC+ testing record). The FFP skill's Step 4 now inlines the correct endpoint, parameter name, and JSON path explicitly: no substrate lookup required.
2. **24x7 shift coverage math missing.** Workers reconstructed from first principles with varying results. Added Step 0.5 "Shift Coverage Staffing" with 4.2 FTE single-seat and 8.4 FTE double-seat formulas and worked example.
3. **Physical engineering SOC's absent from the mapping table.** Workers defaulted Mechanical Engineer, Electrical Engineer, etc. to 15-1211 (Computer Systems Analyst) or 17-2199 (Engineers, All Other). Added explicit 17-2xxx block: 17-2011 Aerospace, 17-2031 Biomedical, 17-2041 Chemical, 17-2051 Civil, 17-2071 Electrical, 17-2072 Electronics, 17-2081 Environmental, 17-2112 Industrial, 17-2141 Mechanical, 17-2161 Nuclear, 17-2171 Petroleum.
4. **PM SOC mapping conflated.** Workers defaulted Program Manager to 11-3021 (Computer and Information Systems Managers) regardless of context. Patched to context-dependent: 11-1021 General and Operations Manager (default / ops), 11-9041 Architectural and Engineering Manager (physical engineering programs), 11-3021 Computer and Information Systems Managers (IT programs only).
5. **Cleveland MSA renumbering not flagged.** Opus S2 burned significant time brute-force scanning. Patched in the downstream BLS skill (Round 3) and cross-referenced in FFP Step 2 with a silent-wrong-answer trap entry.
6. **BLS series ID component lengths not documented.** Workers constructed invalid 24- or 26-char IDs and retried. Added component breakdown: prefix(4) + area(7) + industry(6) + SOC(6) + datatype(2) = 25 chars total. Documented in Step 2 with a worked example.
7. **Seniority modeling absent.** Default wage pull was mean or median only. For Senior/Junior LCATs, workers needed interquartile context. Added P25 → Junior, P50 → Mid, P75 → Senior pattern in Step 2 with explicit instruction to pull all 5 percentiles.
8. **Aging factor hardcoded rather than cell-referenced.** Wave 1 workers applied aging as "× 1.023" hardcoded in formulas. If user changes the contract start assumption, the whole sheet recomputes wrong. Patched Step 2B + Step 8 to require cell-referenced formula:  $=BLS\_2024\_wage * ((1 + escalation)^{months\_gap\_12})$  with BLS\_vintage, contract\_start, months\_gap, and aging\_factor as named assumption block rows (9-12).
9. **Rate validation flag band miscalibrated.** Pre-patch threshold was 10%; legitimate FFP premiums routinely exceeded that. Patched: 0-15% expected, 15-40% FFP premium band, >40% needs justification.

10. **CALC+ dual-pool analysis undocumented.** For senior LCATs, title-match alone often returns  $N < 10$ . Added dual-pool method in Step 4: title-match primary, experience-match secondary, report both counts and both medians.
11. **o-night day trip edge case missing.** Day trips (same-day return) use partial M&IE only, no lodging. Pre-patch Step 5 didn't distinguish. Added explicit o-night case: 75% M&IE first day, no lodging, no last-day M&IE.
12. **"No travel" Sheet 5 handling absent.** Workbook always built Sheet 5 with zeros and placeholder text that broke downstream formulas. Patched: if travel = 0, Sheet 5 says "Travel Not Applicable" with no SUM references.
13. **Multi-location with explicit headcount triggered an unneeded prompt.** Workers asked "Option A (blend), B (lead location), or C (separate lines)" even when user gave per-location headcount. Patched: Option C default when headcount per location is explicit. Prompt only if blend is ambiguous.
14. **FFP-by-deliverable hour allocation method was user-choice with no guidance.** Workers picked differently across 4 S4 runs (uniform split, per-LCAT matrix, staffing-profile weighted). All three are valid; the skill didn't say so. Patched Step 7 with three methods documented, selection guidance by project size, and requirement for worker to cite which method they chose.
15. **Deliverable-timing escalation inconsistent.** Workers sometimes applied escalation within a single PoP, sometimes not. Patched: single-period PoP gets aging-to-start only, no mid-contract escalation. Multi-year PoP applies escalation to out-years per Step 7.
16. **Sheet 2 block layout formulas absent.** Workers built row references by hand for each LCAT block. Patched Step 8 with explicit formula:  $\text{row}(N) = 1 + (N-1) * 19$ , FBR at offset +17, multiplier at +18. Verifiable in a glance.
17. **No explicit final-step "present the file."** Workers wrote to /mnt/user-data/outputs/ but sometimes didn't call `present_files()`. File existed in sandbox but wasn't surfaced to UI. Added Step 9: explicit copy-and-present pattern.

Bonus patches shipped alongside: - Annotation text cannot start with = + - @ (Excel formula parse). Documented in Step 8 with escape guidance. - ODC placeholders must be numeric 0 (not text "TBD") to prevent #VALUE! propagating through SUM formulas. Documented in Step 5. - Implied multiplier column handling when user doesn't want the audit column: drop or annotate as non-billable. - Domain triage first: the skill now instructs the worker to identify agency domain (DoD / IC / DOE / civilian IT / research) before SOC mapping. Domain signals which SOC block applies.

## Wave 2 results (post-patch)

Scenario	Opus 4.7
S1 DC dev team FFP-by-period	14/14
S2 Cleveland 24x7 SOC	14/14
S3 Oak Ridge DOE multi-LCAT	14/14
S4 DISA FFP-by-deliverable	14/14
<b>Total</b>	<b>56/56 (100%)</b>

**Wave 2 aggregate: 56/56 (100%). All 17 Wave 1 issues fixed; no new failures observed.**

## Methodology upgrades observed beyond the matrix

Wave 2 Opus workers produced stronger output even on assertions that passed in Wave 1:

- **S1:** used P25/P50/P75 for Junior/Mid/Senior variants explicitly; cited the patched FFP premium band (15-40%) in rate validation narrative; dropped implied-multiplier column with justification note.
- **S2:** used Cleveland 0017410 directly (no brute-force scan); computed 4.2 FTE via the Step 0.5 worked example; noted the 2024 OMB renumbering explicitly in methodology.
- **S3:** used 17-2141 Mechanical, 17-2071 Electrical, 17-2041 Chemical from the new engineering block; selected 11-9041 Engineering Manager as PM SOC with context justification; applied dual-pool CALC+ for senior engineers with title-match N=4 + experience-match N=27 both reported.
- **S4:** chose staffing-profile allocation with explicit rationale (matrix too complex for 6 LCATs × 4 deliverables, uniform split violated known back-loading of D3+D4); applied aging-to-start only (no mid-contract escalation on 18-month single PoP); cited /v3/api/ceilingrates/ + keyword= explicitly in CALC+ validation methodology; called `present_files()` in Step 9.

## What was not tested

- Sonnet 4.6 on the post-patch skill (inferred from Wave 1 parity on 3/4 scenarios; not directly validated)
- FFP with award fee overlay (hybrid FFP + award fee structures)

- CR-to-FFP conversion modeling (pricing legacy cost-plus scopes as FFP)
- OCONUS travel CLINs (State Department rates)
- ANSI/EIA-748 EVMS-compliant cost buildup formatting
- DCAA forward-pricing rate proposal audits (distinct activity from IGCE build)
- Indefinite Delivery vehicles with seed FFP task orders (ordering-vehicle-level pricing)
- Contract bundling or consolidation scenarios with cross-location overhead pools
- International labor / EU wage data (BLS is US-only)
- Uncertainty quantification beyond the three-scenario band (Monte Carlo, sensitivity analysis)

Wave 3 retest (post-Round 4 substrate validation)

Three Opus scenarios re-run against the patched skill in April 2026, same 42-assertion matrix as Wave 2. Scenarios re-exercised Dayton MSA renumbering (now BLS-patched), DoE M&O overhead environment, GSA MAS commercial burden preset, o-night day trip, and the CALC+ dual-pool pattern.

Scenario	Wave 2	Wave 3 retest
S1 Wright-Patterson DoD Secret engineering	13/14	14/14
S2 Oak Ridge DOE FFP-by-deliverable	13/14	14/14
S3 NASA Glenn GSA MAS 24x7 SOC	13/14	14/14
Total	39/42 (93%)	42/42 (100%)

All three previously-failed burden-band assertions flipped to PASS after Round 5 patches shipped. Zero regressions on the 39 previously-passing assertions.

Round 5 patches shipped (between Wave 2 and Wave 3 retest)

1. **Wrap rate presets by contract vehicle** (10-row table) added to Information to Collect.  
Explicit instruction to ASK about contract vehicle before defaulting to skill mid. Covers GSA MAS commercial/cleared, Agency BPA non-cleared/cleared, DoD prime non-cleared/Secret/SCIF, DoE M&O/FFRDC, R&D CR, OCONUS.

2. **DATEDIF formula fix.** Replaced malformed `YEAR(LEFT(B9,4))` (Excel can't apply YEAR to a string) with `(VALUE(LEFT(B10,4))-VALUE(LEFT(B9,4)))*12 + . . .`. Every Wave 2 worker had to patch this in-place.
3. **"Wait - 19 rows" drafting artifact removed.** Sheet 2 block layout now reads clean.
4. **Cap decision tree extended for P75-also-capped case.** Knoxville Nuclear Sr / LANL physicist pattern. Skill now prescribes: use Mean when P75 caps, cross-reference commercial surveys, apply national P75/median ratio if deriving.
5. **Contract start date default.** Auto-default to October 1 of next federal fiscal year, surfaced as blue-font editable cell. No more silent invention.

## Round 6 patches shipped (Wave 3 retest findings)

Independent workers in Wave 3 each caught the same math error in the Round 5 preset table's "Implied multiplier" column. Three workers independently computed the compounded arithmetic and flagged it.

1. **Preset multiplier column math corrected** across all 10 vehicle rows. Example: GSA MAS commercial 30/60/10/8 was "~1.9x" (wrong); actual math  $1.30 \times 1.60 \times 1.10 \times 1.08 = 2.47x$ . Corrected row-by-row. Added an explicit "Math check" line showing the compounding formula so builders can verify.
2. **Vehicle-aware sanity band.** The generic 2.2x-3.5x commercial band misfires against cleared DoD and DoE M&O builds. Round 6 adds per-vehicle expected ranges: GSA MAS commercial 2.2-2.6x, DoD Secret non-SCIF 3.1-3.4x, DoE M&O 3.0-3.8x, etc. Flag for review only if MID falls outside its vehicle-specific band.
3. **Sheet 5 day-trip IF branch.** Without `IF(B7=0, . . .)` on rows 8 and 10, a day trip (Nights=0) silently produces 150% M&IE instead of the 75% single-partial-day per FTR 301-11.101. This is a workbook-level silent-wrong-answer bug. Template now shows the IF branch explicitly.
4. **CALC+ discovery path added to JSON-paths block.** `aggregations.labor_category.buckets` with `key/doc_count` per bucket. Sits alongside the existing `wage_stats` and `histogram_percentiles` paths. Wave 3 S2 worker had to probe the raw response because this wasn't documented.
5. **Text-starting-with-equals promoted to top-level silent-wrong-answer trap.** Previously buried under Sheet 2 formatting notes. Any cell starting with `=`, `+`, `-`, or `@` is parsed as a formula by Excel, applies to all sheets including Methodology prose.

6. **Cross-sheet DL hourly reference index called out explicitly.** Wave 3 S1 worker hit a \$16.9B fantasy total by indexing off row 4 (Aged Annual Wage) instead of row 5 (DL Hourly). Previously only the FBR index ( $18+i*19$ ) was called out.

## Round 7 patches queued (not shipped)

None block current ship state.

1. Named ranges instead of row-indexed cell references to eliminate row-drift fragility when title banners or extra preamble rows are added to Sheet 1.
2. Mandatory Step 8.5 "Run recalc and verify" rather than parenthetical inside Step 8.
3. Arithmetic consistency check before save (pick one LCAT, one scenario, verify  $\text{FBR} \times \text{hours} \times \text{headcount}$  equals Summary row).
4. DoD cleared engineering worked example in Quick Start (exact Wave 3 S1 pattern).
5. Thin-corpus CALC+ labeling rule: below ~25 records, label as "indicative only, not statistical validation."
6. FFP-by-deliverable Structure B scaffolding expanded to match Structure A depth (CLIN column template, per-LCAT vs uniform formulas, worked example).
7. RSE rubric propagated from BLS skill into FFP Methodology guidance ( $< 5\%$  defensible, 5-15% cite with range,  $> 15\%$  directional only).
8. Adapt FFP workflow patterns into IGCE Builder CR and IGCE Builder LH/T&M skills (already done for 17 cross-cutting patches; Round 5/6 additions not yet ported).
9. SOW decomposition Workflow A+ structured edit gate (add LCAT / rename LCAT / remove LCAT / split) before Step 1.
10. Sheet 2 block size constant cell for future-proofing if block row count changes.

## Wave 4: Post-MCP Migration (Claude Code CLI, Opus 4.7)

### Context

Between Wave 3 and Wave 4, the skill substrate was migrated from three Python L1 skills (bls-oews-api, gsa-calc-ceilingrates, gsa-perdiem-rates calling public APIs) to three dedicated MCP servers (bls-oews, gsa-calc, gsa-perdiem). The MCPs absorb API-key handling, URL construction, series ID assembly, MSA renumbering lookups, JSON path parsing, and the 75% first/last day M&IE rule.

The FFP skill was reduced from 702 to 649 lines at migration by stripping defensive text that the MCPs now obviate (CALC+ q= vs keyword= trap, aggregations.wage\_stats JSON-path archaeology, 25-char BLS series ID assembly, manual 75% M&IE math). Wave 4 tested the reduced skill against the same substrate-free scenarios.

All three rounds ran in Claude Code CLI on Claude Desktop, against the local `~/.claude/skills/igce-builder-ffp/SKILL.md`. An earlier attempt on Claude Desktop chat surfaced a 4-minute hang on the bls-oews MCP's `detect_latest_year` probe; the same probe returned in milliseconds from Claude Code on the same machine, localizing the bug to Claude Desktop's MCP client rather than the server. All 8 federal MCPs were verified healthy from Claude Code.

## Methodology

Each round ran a single scenario in a fresh Claude Code conversation. The worker built the workbook end-to-end. An independent Opus 4.7 grader (separate Claude Code session) then reviewed the produced workbook, read the SKILL.md, and reported findings covering both skill defects and execution gaps. Fixes were applied to SKILL.md between rounds.

Scenarios were chosen to escalate from baseline to judgment-heavy: - **R1:** S1 DC dev team FFP-by-period (3 FTE, GSA MAS commercial, base + 2 OY, no travel) - **R2:** S2 Cleveland 24x7 SOC (single-seat shift coverage, Agency BPA cleared, base + 2 OY, quarterly DC travel) - **R3:** S3 Oak Ridge DOE 18-month feasibility study (6 LCATs, FFP-by-deliverable 15/30/25/30, DoE M&O, no travel)

## Round 1 findings (S1 DC dev team, GSA MAS commercial)

Workbook built cleanly. Mid total ~\$3.10M, implied multiplier 2.47x, zero formula errors. Six skill issues surfaced:

1. **BLS datatype list stale.** Skill requested [04, 11, 12, 13, 14, 15, 02, 05]. MCP rejected 02 and 05 (employment and wage RSE). Valid set is 01, 03, 04, 08, 11, 12, 13, 14, 15.
2. **Step 9 claude.ai-specific.** Hardcoded `/mnt/user-data/outputs/` and `present_files` neither of which exist on Claude Code CLI.
3. **Rate validation >40% threshold over-triggers.** Skill's narrative said 15-40% typical for DC/high-cost metros, but the formula flagged anything above 40% as "requires justification." DC Software Developer mid FBR landed at 57% above CALC+ P50 and got flagged despite the skill's own calibration note.



4. **No default for "N-person team" without seniority tiers.** Skill documented junior/mid/senior percentile conventions but silent on how to price a generic "3-person team."
5. **Preset vs generic wrap-rate table hierarchy ambiguous.** Generic Low/Mid/High table showed Mid =  $32/80/12/10 = 2.93x$ , which matches DoD non-cleared preset, not the GSA MAS commercial preset (2.47x). Worker could read either as authoritative.
6. **Platform-level: evaluator reported \$Bword substitution tokens** in the Sheet 2 block layout ("\$Bdevelopment", "\$Byears", "\$Bteam"). Grep of source file returned zero matches. Not in skill; Claude Code skill-loader substitution artifact OR evaluator hallucination.

**Fixes shipped:** corrected datatype list, environment-aware Step 9 with CLI fallback, rate band recalibration (0-15 / 15-40 / 40-70 / >70 with explicit DC/metro premium band), N-person P50 default, relabeled generic table as "sensitivity reference only," simplified block-layout guardrail to remove confusing example tokens.

## Round 2 findings (S2 Cleveland 24x7 SOC, Agency BPA cleared + travel)

Workbook built cleanly. Mid 3-year total ~\$4.24M. Seven skill issues surfaced:

1. **Five wrong multipliers in Vehicle Preset table.** Stated vs actual: GSA MAS cleared 2.59→**2.87**, Agency BPA non-cleared 2.53→**2.85**, Agency BPA cleared 2.91→**3.17**, DoD SCIF 3.67→**3.64**, R&D BAA CR 2.99→**3.03**. Worker caught the 3.17x Agency BPA cleared discrepancy during the build and documented the corrected multiplier in Methodology, but a less careful operator would have shipped the stated 2.91x.
2. **Sanity bands excluded actual preset values.** The stated 2.8-3.0x band for Agency BPA cleared / DoD non-cleared excluded the true 3.17x Agency BPA cleared multiplier.
3. **Shift-coverage travel ambiguous.** Step 0.5 derived 4.2 FTE for single-seat 24x7 but said nothing about how many travel per trip. Worker picked 1 (shift-lead rotation) and noted the assumption.
4. **SOC 15-1212 InfoSec Analyst fragile at metro level.** Cleveland MSA 17410 suppressed for this SOC; worker fell back to Ohio state. Skill mentions generic metro→state→national fallback but doesn't flag InfoSec Analyst as a known-fragile SOC (common in most mid-size metros outside tech hubs).
5. **CALC+ "SOC Analyst" query fragmented.** 33 records spread across 27 buckets, max bucket 2 records. Useful pool was "Information Security Analyst II" (31 records). Skill had no canonical-query hint for this common LCAT term.
6. **FY2027 per diem fallback worked cleanly.** MCP returned empty rates array for FY2027; worker fell back to FY2026 per skill rule. No change needed; flagged as skill strength.



7. **Platform-level:** evaluator again reported \$Banalyst, \$B0Ys, \$Bcoverage substitution tokens. Same non-issue as Round 1.

**Fixes shipped:** corrected all 5 multipliers to match actual arithmetic, recalibrated sanity bands, added "1 representative per trip" default for shift-coverage travel, added known-fragile SOC note to Step 2, added canonical CALC+ query hints for fragmented LCATs ("SOC analyst" → Information Security Analyst I/II/III).

## Round 3 findings (S3 Oak Ridge DOE 18-month feasibility, 6 LCATs, FFP-by-deliverable)

Workbook built cleanly. Mid total ~\$2.94M, DoE M&O multiplier 3.18x confirmed. Seven skill issues surfaced:

1. **B8 "Base Year Months" doesn't fit single-period PoPs.** Worker renamed to "Period Months" = 18 and adjusted hours formula. Skill silent on this pattern.
2. **No-travel Sheet 1 row not explicit.** Skill says build Sheet 5 as "Not Applicable" but said nothing about the Summary. Worker added \$O Travel row on Sheet 1 with "TBD" note.
3. **Sanity band not pinned to preset row.** DoE M&O 3.18x was within its band but worker had to cross-reference two separate tables (preset + band) to confirm.
4. **Methodology formula-ref rule too narrow.** Rule only explicitly applied to aging factor. Worker hardcoded "1.0615" and "2.47x" as text strings in narrative; those go stale if user edits B6 or B10.
5. **No post-build sanity check.** Row 4 vs row 5 DL hourly reference trap remains a silent \$B-dimension bug; skill listed it as a silent-wrong-answer trap but no mandatory validation step.
6. **Nuclear Engineer distribution compressed at Knoxville.** P25 = P10 = \$93,980 (ORNL/Y-12 concentration crushes the lower half). Skill cap decision tree covered P75 caps but not P25=P10 compression.
7. **Platform-level:** third consecutive round reporting substitution tokens (\$Bfeasibility, \$Bscope). Source file clean on grep. Locked in as a platform-layer artifact.

**Fixes shipped:** relabeled B8 as "Base Year Months (or PoP Months)" with inline rename guidance, explicit Sheet 1 no-travel row instruction, pinned Expected band column to each Preset table row, mandatory Methodology formula-ref rule at top of Sheet 6 spec, new Step 8.5 post-build sanity check with dimensional  $\text{avg\_FBR} \times \text{hours} \times \text{FTE}$  guardrail, extended cap decision tree with compressed-distribution branch.

## Wave 4 aggregate

Metric	Value
Rounds	3
Workbooks produced	3
Workbooks that opened without #VALUE! errors	3
Workbooks with implied multiplier matching vehicle preset	3
Silent-wrong-answer bugs observed	0
Skill defects identified by evaluator	18
Skill defects fixed between rounds	18
Platform-layer substitution reports (not skill bugs)	3
Line delta: SKILL.md post-Wave-3 (702) → post-Wave-4-fixes	689

### Platform-level finding (not actionable in the skill)

Three consecutive rounds reported \$B<prompt-word> substitution tokens in the Step 8 Sheet 2 block layout (e.g., \$Bdevelopment, \$Byears, \$Banalyst, \$BOYs, \$Bfeasibility, \$Bscope). Grep of the SKILL.md source returned zero matches each time; the file contains literal integer cell addresses (\$B\$2, \$B\$12, etc.). Two plausible root causes:

- **Claude Code skill-loader substitution:** the loader may apply a template-style substitution on \$VAR-shaped tokens in the skill markdown before handing it to the model.
- **Evaluator model hallucination:** the evaluator reads the cell addresses correctly but, when describing the substitution failure mode warned about in the guardrail, confabulates examples that match the pattern.

Either way, the skill cannot fix this at the source. Mitigation: removed example tokens from the guardrail to reduce the evaluator's priming surface. Root-causing requires instrumenting the skill loader, which is out of scope for Wave 4.

## What has not been tested in Wave 4

- **Workflow A+ SOW/PWS decomposition gate.** Raw SOW text input, Step 0 validation gate, user confirmation before Steps 1+.
- **Multi-location with explicit headcount.** Option C separate-lines path; two metros with defined FTE splits and inter-site travel.
- **Workflow B rate validation only.** `mcp__gsa-calc__price_reasonableness_check` shortcut, dual-pool analysis, no workbook.
- **Cap stress at multi-capped metros.** LANL/SF/NYC where P90 and P75 both cap for specialty occupations.
- **Custom wrap rate workflow.** CO provides explicit rates (cleared/SCIF); worker applies as MID, generates LOW/HIGH offsets.
- **Partial base year.** Mid-year award start with 9-month base period.
- **Sonnet 4.6 parity on Wave 4 fixes.** All three rounds were Opus 4.7.

These are queued for Wave 5.

## Fixes shipped cumulatively in Wave 4

All 18 fixes ship together in the current `SKILL.md` at the head of this testing record. Chronological order (Round 1 → Round 3):

1. BLS datatype list updated (dropped invalid 02/05)
2. Step 9 environment-aware with Claude Code CLI fallback
3. Rate validation bands recalibrated (0-15 / 15-40 / 40-70 / >70) with metro premium acknowledged
4. N-person team no-tiers default (all at P50)
5. Block-layout guardrail simplified
6. Generic Low/Mid/High table relabeled "sensitivity reference"
7. 5 Vehicle Preset multiplier arithmetic corrections
8. Sanity bands pinned to each preset row (new column)
9. Shift-coverage travel default (1 representative per trip)
10. Known-fragile SOC note (15-1212 InfoSec Analyst, 15-2051 Data Scientist, 19-2012 Physicists)
11. Canonical CALC+ query hints for fragmented LCATs
12. B8 relabeled for single-period PoPs
13. Sheet 1 no-travel row made explicit

14. Methodology formula-reference rule promoted to mandatory top-of-section
15. Step 8.5 post-build sanity check (dimensional guardrail)
16. Cap decision tree extended for compressed P25=P10 distributions
17. Vehicle preset "Notes" column reshaped for clarity
18. "Sanity band is vehicle-aware" paragraph consolidated (info moved into preset table)

## Wave 5: ai-boundaries + untested workflow paths (Claude Code CLI, Opus 4.7)

### Context

Wave 4 covered the three main Workflow A build paths (FFP-by-period single-location, shift coverage + travel, FFP-by-deliverable multi-LCAT). Wave 5 covered the remaining untested paths: Workflow A+ SOW decomposition gate (S4), multi-location with explicit headcount (S5), and Workflow B rate validation (S6). Wave 5 also applied the repository's ai-boundaries.md as a grading lens for the first time, which surfaced a Tier-1 violation in Workflow B.

### Round 1 findings (S4 Fort Meade DoD BPA cybersecurity PWS, Workflow A+)

Workbook built ~\$21.4M mid total. Worker paused at the Step 0 validation gate and used AskUserQuestion before building - partial pass on the gate (see below). Seven findings:

1. **Step 0 validation gate conflates decomposition with build parameters.** Worker presented the decomposition table and then immediately asked parameter questions (wrap preset, shift coverage, metro) in the same AskUserQuestion call. User rubber-stamped the decomposition by answering parameter questions; no explicit decomposition approval. **Fix: separate Stage A (decomposition OK?) from Stage B (parameters).**
2. **Missing preset: DoD BPA + SCIF stack.** Presets force either-or between Agency BPA cleared (3.17x) and DoD SCIF (3.64x); a BPA operating inside a SCIF (common at Fort Meade / NSA) doesn't fit either cleanly. **Fix: added DoD BPA (TS/SCI SCIF) row at 32/115/13/10 → 3.39x with 3.2-3.6x band.**
3. **Per-block Sheet 2 formulas block-1-indexed.** Step 8 shows  $B=B5*B7$ ,  $B=B12+B14$  as if every block were block 1 (row 1). Block 2 starts at row 20; a builder copying formulas verbatim into block 2 references block 1 cells. **Fix: explicit base\_row = 1 + (N-1) \* 19 with worked example for block 2 formula shifts.**

4. **Stacked-premium worked example missing.** CALC+ divergence bands (15-40, 40-70, >70) don't explain how stacked factors produce large percentages. Worker's InfoSec FBR landed 100-140% above national CALC+ P50 because metro × P75 × aging × SCIF/commercial wrap ratio all stacked. **Fix: Step 4 now includes a worked decomposition table (metro × seniority × aging × wrap-ratio = expected ratio).**
5. **Tier 1 vs Tier 2 distinction in Step 0.5.** Skill treats "24x7 Tier 2" identically to "24x7 SOC." In practice Tier 2 is often an on-call overlay on Tier 1 (2-3 FTE) rather than a 4.2 FTE layer. **Fix: added clarifying question trigger via AskUserQuestion when PWS says "Tier 2" specifically.**
6. **TS/SCI compliance overhead not flagged.** NIST 800-53 continuous monitoring, STIG remediation, accreditation maintenance run 5-10% of staff time on cleared contracts. Not in BLS wages. **Fix: added optional 5-10% buffer option with user-confirmation gate.**
7. **Minor ai-boundaries observation.** Worker described rate divergence as "Defensible but will draw reviewer questions" - model-originated "defensible" conclusion. **Fix: rolled into the Tier-1 ai-boundaries scrub.**

Platform-layer: evaluator reported \$BFFP, \$BIGCE, \$Bpriced substitution tokens. Source file clean on grep. Fourth consecutive round. Locked in as Claude Code skill-loader substitution artifact.

## Round 2 findings (S5 Fort Meade + Colorado Springs multi-location, Workflow A)

Workbook built cleanly. ~\$9.0M mid 3-year total. Worker correctly used Option C (separate lines per location) without prompting, handled FY2027 per diem fallback cleanly, ran Python-side dimensional sanity check (no LibreOffice). Six findings:

1. **DoD installation → GSA per diem city crosswalk gap.** "Fort Meade" returned empty; GSA keys it under Annapolis / Anne Arundel County. Single most common friction point for DoD users. **Fix: added crosswalk table to Step 5 covering 15 high-traffic installations (Fort Meade, Belvoir, Pentagon, Liberty, Peterson, Wright-Pat, Eglin, NSA Bethesda, etc.).**
2. **"Travel between sites" ambiguity.** "Quarterly travel between sites" with 2+ destinations could mean 4 total or 4 each way. Worker defaulted to total-split-evenly and flagged. **Fix: canonical rule added: trips/year TOTAL split evenly unless user says "each way" explicitly.**
3. **Per diem FY fallback trigger mismatch.** Skill said "fallback if MCP returns empty array"; MCP actually returns an explicit error string for unpublished FYs. **Fix: fallback now triggers on both empty array OR error containing "No rates found for FY{year}".**

4. **Multi-destination Sheet 5 layout parameterization missing.** Step 8 Sheet 5 shows one per-destination block; for M destinations builders invent the block-2 starting row. **Fix: added block N at row 1 + (N-1) \* 17 parameterization and cross-sheet SUM formula template.**
5. **CLI recalc fallback gap.** Step 8.5 and Step 9 assumed `/mnt/skills/public/xlsx/scripts/recalc.py` exists. Not available on Claude Code CLI without LibreOffice. Worker used parallel Python computation against raw inputs. **Fix: Step 8.5 now explicitly handles three environments (claude.ai web, Claude Code CLI, macOS Numbers).**
6. **Methodology formula-ref rule underweighted.** Rule buried mid-Step 8. Worker violated it by hardcoding "3.2525x" as string. **Fix: rule promoted to mandatory callout at top of Sheet 6 spec in Wave 4; reinforced in Wave 5 with concrete =TEXT() patterns.**

Platform-layer: evaluator did not report substitution tokens this round (or did not emphasize). Pattern stays locked.

### Round 3 findings (S6 Senior Data Scientist \$225/hr DC Agency BPA, Workflow B) - ai-boundaries violation

Worker produced a Price Reasonableness Determination memo and declared the rate "fair and reasonable" in Section 7. Under ai-boundaries.md grading, this is a **Tier-1 violation**:

1. **Model-originated "fair and reasonable" conclusion.** Worker opened the chat response with "Yes, reasonable" and the memo Section 7 asserted "determined to be fair and reasonable in accordance with FAR 15.404-1(b)(2)(i) and (v)." The CO's determination was written by the model. ai-boundaries.md Rule 2: "If the signer cannot defend every evaluative claim in the final record without pointing back at the tool's output, the tool crossed the line."
2. **Invented TS/SCI clearance premium.** 15-25% premium applied in memo Section 5 as a market fact. Worker's own evaluator notes acknowledged: "It's not in the skill. Agents supply it from general knowledge." Model-originated rationale treated as data.
3. **Advisory text in chat.** "I'd push back only if the vendor can't articulate the clearance value..." - model telling the CO how to negotiate.

**Fixes shipped: - Workflow B rewrite.** From "Position each rate: below 25th (aggressive), 25th-75th (competitive), above 75th (premium), above 90th (outlier requiring justification)" to "Pull data and describe positioning neutrally; do NOT assert fair/reasonable/defensible." Calibration band labels for Sheet 4 Status column rewritten as positional: "Within CALC+ FFP premium range" / "Metro geographic premium; see Methodology for factor decomposition" / "CO review recommended for factors outside BLS/CALC+ data." - **Memo drafting gate.** Skill now drafts a price reason-

ableness memo ONLY when the CO supplies the rationale and conclusion in the prompt; memo template leaves Determination section as [CO to complete] placeholder unless the CO supplied it. Skill responds to naked "draft the memo" requests with: "Provide the rationale you want documented and your fair-and-reasonable conclusion; I'll format it." - **Out-of-data premiums named as gaps.** TS/SCI premium, OCONUS hazard, SCIF overhead, specialty labor market: if not in BLS/CALC+/Per Diem data, skill flags the gap. No model-originated premium ranges. - **ai-boundaries citation.** New "Operating Principle (ai-boundaries)" section at the top of the skill names the rule explicitly with examples of what the skill does and does not do. - **Evaluative-verb scrub.** "Defensible," "reasonable," "acceptable," "competitive," "outlier" removed from narrative-generation paths (Methodology sheet prose, chat summary, validation sheet status).

Wave 5 aggregate

Metric	Value
Rounds	3
Workbooks / documents produced	2 workbooks + 1 memo
Tier-1 ai-boundaries violations identified	1 (S6)
Skill defects identified total	22
Skill defects fixed	22
Platform-layer substitution reports	4th confirmation (S4); S5/S6 did not emphasize

Pre-flight MCP check added

Separate from findings, a new pre-flight block was added at the top of the skill to verify the three MCPs (bls-oews, gsa-calc, gsa-perdiem) are available and their API keys are configured before any workflow executes. Missing MCP: stop, tell the user which MCPs are missing, ask them to install and configure. Missing API key: stop, tell the user which key is missing. The skill does not attempt to work around missing MCPs by calling APIs directly.

Platform-level finding (cemented)

Four consecutive rounds across Waves 4 and 5 have reported \$B<prompt-word> substitution tokens in the Step 8 block layout (examples: \$Bdevelopment, \$Byears, \$Banalyst, \$BOYs,



\$Bfeasibility, \$Bscope, \$BFFP, \$BIGCE, \$Bpriced, \$Bper, \$Bof). Every token maps to a word from the session's user prompt. Grep of SKILL.md source returns zero matches every time. Root cause is in the Claude Code skill-loader substitution pipeline OR a consistent evaluator hallucination pattern keyed off the block-layout template. Not fixable in the skill. Mitigation in place: example-token-bearing guardrail text was removed; literal \$B\$2 / \$B\$12 notation retained.

## What has not been tested

- **Custom wrap rate workflow.** CO provides explicit rates (cleared/SCIF/OCONUS); worker applies as MID, generates LOW/HIGH offsets. The skill has the rule; no test run has exercised it.
- **Cap stress at multi-capped metros.** LANL / SF / NYC where P90 AND P75 both cap for specialty occupations; tests the Step 2 cap decision tree including the compressed-distribution branch.
- **Partial base year proration.** Mid-year award with 7-9 month base period (e.g., March start against Sep 30 FY end).
- **Memo drafting with CO-supplied rationale.** Wave 5 surfaced the need for the memo gate; the gate itself is untested.
- **Sonnet 4.6 parity across Waves 4 and 5.** All runs were Opus 4.7.

These are queued for Wave 6.

## Wave 6: Cross-skill findings port (Claude Code CLI + Desktop, Opus 4.7)

### Context

Wave 6 did not run FFP-specific scenarios. Instead, it ported horizontal findings discovered during LH/T&M Wave 2 testing plus shipped a hardened v2 ai-boundaries gate to replace the Wave 5 patch that proved insufficient in live testing.

The v2 ai-boundaries gate was forced by a live LH/T&M test where the Wave 5 patch failed. The LH/T&M skill (carrying the same Wave 5 ai-boundaries language as FFP) drafted a full price reasonableness memo with 5 separate "rate is fair and reasonable" determinations, recommended negotiation positions toward CALC+ P75, and drafted Evaluation Notice language, all forbidden by the Wave 5 ai-boundaries patch. Root cause: Wave 5 placed the gate at Workflow B Step 6 "Stop" which is too far downstream; by that point the model was committed to helpful-memo-author momentum and the "Stop" instruction read as advisory rather than blocking. Fix: moved the gate to Step 0 with a to-



ken-scan + verbatim refusal template + Option A/B bifurcation (Option A = positioning data only; Option B = memo template fill with CO's verbatim rationale and determination).

FFP Workflow B was updated with the same v2 gate.

## Horizontal findings ported from LH/T&M Wave 2

Six patches were ported to FFP:

1. **CALC+ keyword\_search** → **igce\_benchmark redirect** for stats-only queries.  
igce\_benchmark returns percentiles without the full record list; faster and avoids false signals on large corpora.
2. **Tier-matched keyword rule.** Query each seniority tier (P25/P50/P75) with its own keyword string, not the aggregate pool. Avoids false divergence flags when a Senior LCAT compared against an aggregate title-match pool reads as overpriced because the pool contains Juniors.
3. **NSA Bethesda per diem crosswalk fix.** DoD installation crosswalk pointed NSA Bethesda at Montgomery County. NSA Bethesda staff living in Bethesda use the DC composite locality, not Montgomery County, per GSA convention. Crosswalk table updated.
4. **FY rollover guidance.** If contract PoP start is within 6 months of next FY, query both FYs and document refresh-on-publication. Avoids locking workbook into soon-to-expire rates.
5. **Raw Data sheet granularity rule.** Use summary tables with query parameters inline, not raw JSON dumps. Several Wave 2 LH/T&M workbooks shipped with 40KB+ of raw CALC+ bucket JSON that added nothing readable to the audit trail.
6. **Step 9 CLI branch.** present\_files is claude.ai-only. CLI path is simple file-write-then-report. MacOS Desktop with Numbers is third branch. Wave 5 already partially covered this; v2 consolidates the three-environment fork.

Plus:

1. **Stage A/B skip for structured inputs.** Workflow A with structured handoff (SOW/PWS builder output) does not need the Stage A decomposition approval; only Workflow A+ from raw SOW text needs it.

## FFP-specific bloat trimmed

The skill had accumulated cruft from prior waves. Trimmed in Wave 6:

- **Known-fragile SOC's paragraph** (added in Wave 4) collapsed into a two-row entry in the Information to Collect table. The paragraph repeated the SOC mapping table's warnings without adding new content.

- **Stacked premium worked example** (added in Wave 5 Round 1) reduced from a full-page decomposition to a 4-row inline table.
- **Quick Start** cut from 12 examples to 4. The 4 retained cover the distinct pricing-structure decision gates (FFP-by-period, FFP-by-deliverable, multi-location, rate-validation-only). The trimmed 8 were restatements of the same decisions against different agencies.
- **Edge Cases** trimmed from a mixed list of genuine traps and quality suggestions down to silent-wrong-answer traps only. Quality suggestions went into a new "Optional enhancements" appendix.

## Line delta

SKILL.md: 897 → 854 (-43 lines).

## Status

All Wave 6 patches were inherited from LH/T&M testing, not directly re-tested on FFP. FFP regression testing against S1-S6 on the post-Wave-6 skill is queued.

## Wave 7 (inherited from CR Wave 1 lazy-prompt testing)

**Wave 7** (Cross-skill patches inherited from CR Wave 1 lazy-prompt testing): CR Wave 1 surfaced 22 findings across three lazy-prompt scenarios. 14 were patched, 8 dropped as too scenario-specific. Universal patches ported to FFP: Installation to GSA locality crosswalk expanded with 6 DOE labs (Oak Ridge, LANL, Hanford, Sandia, LLNL, INL), BLS MSA URL fallback, Workflow A ambiguous-input rule, Step 9 env fork with macOS Excel/Numbers branch, BLS wage-cap 10% proximity rule, shift coverage upfront in Information to Collect, Methodology depth guidance. Editorial fixes: Rate Validation status text neutralized, Sheet 5 travel skip-or-include resolved, Stage A/B skip clarified, CALC+ igce\_benchmark promoted to default, NAICS/PSC proactive ask. **Status:** patches inherited from CR testing, not re-tested on FFP directly. FFP remains validated through Wave 5 end-to-end scenarios plus Wave 6 gate hardening.

## Independent grading methodology

The Wave 1 and Wave 2 testing records were produced under a consistent methodology:

- Scenarios and assertion matrices were committed in writing before any worker output was read
- The grader did not coach workers during runs

- Assertions were graded strict on literal wording; ambiguous assertions were noted and refined for the next wave (not retroactively reinterpreted)
- Methodology is auditable in the `igce-ffp-wave1-runbook.md` and `igce-ffp-wave2-runbook.md` source files
- All findings come from direct observation of worker output and produced workbooks, not inference from memory of prior sessions
- Downstream skill patches (BLS Rounds 2 and 3, CALC+ Round 2) shipped before IGCE FFP Wave 2 so the substrate was validated in the Opus retest

## Wave 8 (universal patches inherited from CR Wave 2 detailed-prompt round)

**Wave 8** (Universal patches inherited from CR Wave 2 detailed-prompt round): 11 universal-principle patches ported to FFP from CR detailed-prompt testing. Includes `page_size=0` update, 24x7 math reconciliation, DATEDIF real-date fix, day-trip M&IE correctness fix (was shipping 25% low), aged-wage row explicit, Sheet 2/Sheet 1 unit clarity, flat-tail detection, 6 DoD/DOE test ranges added to installation crosswalk, SOC 17-2199 fallback, same-metro TDY check, stacked factors definition. Status: inherited, not re-tested on FFP directly.

## Wave 9 (universal patches inherited from CR Wave 4 + LH/T&M Wave 5)

**Wave 9** (Universal patches ported from CR Wave 4 DCAA/FPRA override + LH/T&M Wave 5 Workflow B gate hardening). Two patches shipped to FFP identically to the LH/T&M Wave 5 pair:

1. **CO-supplied DCAA-audited rates override rule** (enhancement to existing Custom rate workflow). Adds explicit language: use FPRA rates as point estimate, do NOT bookend  $\pm 20\%$  around an audited rate (the audited rate IS the rate, not a midpoint), document FPRA effective date and approving authority in Methodology. Trust CO-supplied rate over vehicle-preset band even when they diverge; note divergence in Methodology rather than reconciling to the table. This closes the gap from CR Wave 4 Test 2 (FEMA Booz Allen FPRA) and LH/T&M Wave 5 Test 1 (Lockheed NSA Fort Meade FPRA) which both showed skills treating audited rates as midpoints rather than point estimates.
2. **Workflow B gate fires unconditionally on entry** (bypass fix). Prior gate was token-gated: a prompt like "validate these wrap rates" would route to Workflow B → Step 0 → scan finds none of the listed tokens ("memo," "determination," "fair and reasonable," etc.) → waves through to

Steps 1-5 without presenting the Option A/B refusal template. LH/T&M Wave 5 Test 4 surfaced this as a universal silent-bypass; same path existed in FFP (FFP Workflow B triggers at line 92 include "validate these wrap rates" and "check this FFP proposal," neither of which match the Step 0 token list). Patch makes the gate fire unconditionally on Workflow B entry, with additional hard prohibitions added when memo-drafting tokens also appear (expanded list: "reasonable" standalone, "validate," "acceptable," "justify").

**Status:** both patches inherited from LH/T&M Wave 5. Wave 5 tested DCAA/FPRA on LH/T&M directly and the Workflow B gate on LH/T&M directly; FFP carries the same structural pattern and the patches apply identically. Regression on FFP is deferred to Wave 10.

**Line delta:** 854 → 880 (+26). Ceiling remains 1,000.

*Testing record prepared April 2026 by James Jenrette / 1102tools. Independent grading methodology. MIT licensed. Source: [github.com/1102tools/federal-contracting-skills](https://github.com/1102tools/federal-contracting-skills).*