
TESTING RECORD

IGCE Builder LH/T&M Testing Record

Five waves of testing across April 2026 validated the Labor Hour / Time-and-Materials IGCE skill through 3 end-to-end scenarios plus Wave 5 complete with 3 universal patches including a Gate 3 cell reference correctness fix.

Skill

igce-builder-lh-tm
github.com/1102tools/federal-contracting-skills

Date

April 2026
1102tools.com

Part 1: For Federal Acquisition Users

The bottom line

Independent testing in April 2026 (3 end-to-end runs, 42 binary assertions graded, Claude Opus 4.7) validates the IGCE Builder LH/T&M skill across three real-world federal acquisition scenarios: pure Labor Hour at Redstone Arsenal, T&M hybrid with materials at NAVWAR San Diego, and LH with multi-site travel at AFLCMC Wright-Patterson.

Wave 1 aggregate: 39 of 42 assertions passed (92.9%). Round 1 patches shipped to address all finding categories.

Scenarios tested and how reliably they work

Scenario	Score	Result
S1: Pure Labor Hour at Redstone Arsenal (Huntsville MSA, 4 LCATs, base + 4 options)	14/14	Reliable
S2: T&M hybrid with \$250K/yr materials at NAVWAR San Diego (4 LCATs, base + 2 options)	12/14	Materials handling fee not surfaced (soft fail); FAR 16.601(c)(3) surveillance memo missing
S3: LH with multi-depot travel at AFLCMC Wright-Patterson (Dayton MSA, 4 LCATs, 3 destinations, base + 3 options)	13/14	FAR 31.205-46 travel cost principle not cited

All three scenarios produced workbooks that delivered correct final numbers (once worker caught cell-reference drift mid-build in S2). The failures were gaps in narrative/methodology completeness and skill hardening gaps that allowed those gaps to slip through. All findings patched in Round 1.

Manual-verification checklist

Scan every LH/T&M IGCE output for these before using in a contract file:

- 1. Skill announced itself.** First line of the worker's response should acknowledge "IGCE Builder LH/T&M" was loaded. If the worker started building without naming the skill, the skill likely did not trigger and you are getting a generic xlsx build instead of a hardened IGCE.
- 2. Per-FTE annualized cost is defensible.** Burdened hourly \times productive hours \times 1 FTE should land in \$100K to \$1M. Outside this range (especially above \$1M) indicates formula cell-reference drift.
- 3. Sheet 1 Grand Total equals Sheet 2 Mid-Scenario Grand Total.** Any divergence indicates cross-sheet reference drift.
- 4. BLS aging factor is cell-referenced formula, not hardcoded.** Aging factor must be $= (1+B5)^{(B10/12)}$ or equivalent; changing contract start date should cascade.
- 5. FAR citations complete.** Must see 16.601 always, 16.601(b)(2) for T&M materials, 16.601(c)(2) for LCAT ceiling hours, 16.601(c)(3) for T&M surveillance, 31.205-46 when travel is in scope.
- 6. Raw Data sheet shows rejected SOC alternatives** when a judgment-call re-pick happened (for example, IT PM switched from 15-1299 to 11-3021).
- 7. Materials at cost (no burden) for T&M.** FAR 16.601(b)(2). Handling fee decision explicit in methodology, whether applied or not.
- 8. Travel math uses FTR 301-11.101 75% first/last day M&IE rule** and day trips use single-partial-day M&IE (not full + two 75%).

What the skill does not do

- **It does not produce FFP or CR estimates.** Use IGCE Builder FFP or IGCE Builder CR.
- **It does not produce OT Cost Analyses.** Use OT Cost Analysis skill.
- **It does not substitute for a contracting officer's price reasonableness determination.** IGCE is an estimate; the CO makes the determination per FAR 15.4.
- **It does not guarantee a specific burden multiplier.** Multiplier is a scenario input; user-provided values win over defaults.
- **It has not been tested on:** CONUS-to-OCNUS mixed performance, 24x7 shift coverage with the Step 0.5 math, Workflow A+ SOW/PWS decomposition from scratch, Workflow B rate validation only.

Part 2: For Developers and Technical Reviewers

Testing methodology

Scenarios

Three scenarios designed to exercise distinct LH/T&M mechanics:

- **S1 (LH, baseline, no materials, no travel):** Redstone Arsenal software sustainment, 4 LCATs including a senior tier, GSA MAS IT Schedule 70 vehicle, base + 4 options. Tests burden multiplier math, BLS aging, CALC+ validation, LCAT ceiling hours table, FAR 16.601 citation, escalation.
- **S2 (T&M hybrid, materials at cost):** NAVWAR San Diego network security, 4 cybersecurity LCATs, \$250K/yr materials, 60% on-site + 40% telework, base + 2 options. Tests FAR 16.601(b) (2) materials at cost, FAR 16.601(c)(3) surveillance memo, handling fee decision gate, cybersecurity SOC mapping, materials ceiling language.
- **S3 (LH with multi-depot travel):** AFLCMC Wright-Patterson logistics, 4 LCATs, quarterly site visits to Hill AFB/Tinker AFB/Robins AFB, OASIS+ commercial vehicle, base + 3 options. Tests Dayton MSA 19430 (renumbered from 19380), GSA Per Diem pulls for three destinations, FTR 301-11.101 75% M&IE rule, FY-not-yet-published fallback, FAR 31.205-46 travel cost principle.

Each scenario had a 14-point binary assertion matrix covering skill activation, data source correctness, burden multiplier defensibility, FAR citation completeness, workbook structural integrity, methodology completeness, and staffing handoff respect.

Environment

- claude.ai web chat, fresh conversation per scenario
- Skills installed: `igce-builder-lh-tm` plus required L1s (`bls-oews-api`, `gsa-calc-ceilingrates`, `gsa-perdiem-rates`)
- Model: Claude Opus 4.7 on Max plan
- All three scenarios hit tool-use limits and required one "continue" per run

Grading

Grader (Claude Code session separate from worker runs) read the worker's final response plus the produced xlsx. Workers not coached during runs. Assertions graded binary pass/fail. Partial credits allowed only with explicit notation.

Wave 1 results

Scenario	Score	Fails
S1 Redstone LH	14/14	—
S2 NAVWAR T&M	12/14	Materials handling fee decision not surfaced; FAR 16.601(c)(3) surveillance memo missing
S3 AFLCMC LH+travel	13/14	FAR 31.205-46 not cited for travel
Total	39/42 (92.9%)	3 fails

Round 1 findings: 17 skill bugs surfaced

Across three workers' self-critiques plus direct grader observation, 17 distinct findings emerged. All shipped in Round 1.

Po (must-fix, skill produced wrong numbers or failed to load resources)

1. **Stale CALC+ URL in Step 4 (lines 294, 300).** Skill cited <https://calc.gsa.gov/api/v3/api/ceilingrates/> which returns HTTP 404. Correct URL lives in the CALC+ skill itself. S2 and S3 workers burned round-trips discovering the drift. Patched by removing hardcoded URL and referring to the CALC+ skill as authoritative.
2. **Assumption block cell-reference drift (Step 8).** Skill prose at line 465 did not explicitly map every downstream \$B\$n reference. S2 worker shipped formulas referencing \$B\$4 as escalation (actually Burden High = 2.2) and \$B\$6 as Base Year Months (actually Productive Hours = 1920), producing \$105M-per-FTE base year numbers before catching on value inspection. Recalc

did NOT flag this because formulas were syntactically valid. Patched with explicit DOWNSTREAM CELL REFERENCES block to memorize before writing Sheet 1.

3. **Post-recalc per-FTE sanity gate missing.** `recalc.py` returning zero formula errors is necessary but NOT sufficient; syntactically valid formulas can reference wrong cells and produce wildly wrong values. Patched with Step 8.5 requiring per-FTE cost check in [\$100K, \$1M] range, plus Sheet 1 == Sheet 2 cross-check and burden-multiplier cross-sheet check.

P1 (ship this round, completeness and resilience)

1. **Sheet 1 Grand Total == Sheet 2 Mid-Scenario Grand Total** post-recalc assertion added to Step 8.5.
2. **Sheet 5 merged-cell collision pattern** breaks `openpyxl` ("MergedCell" object attribute 'value' is read-only"). S2 worker hit this and had to refactor. Patched with explicit rule: do NOT merge section-header cells while also writing values to column B of the merged range. Use dedicated header rows.
3. **FAR 31.205-46 (travel costs) not required in methodology.** S3 cited FTR 301-11.101 for M&IE 75% but missed 31.205-46. Patched into required FAR citation set.
4. **FAR 16.601(c)(3) (T&M surveillance) not required in methodology.** S2 missed this. Patched into required FAR citation set with explicit required language.
5. **Materials handling fee decision not surfaced.** S2 applied pure at-cost silently without flagging the handling fee decision. Patched with explicit Materials Handling Fee Decision Gate in Step 5B; default to at-cost but require explicit mention; cite FAR 31.205-26 if fee applied.
6. **BLS 503 retry guidance missing from orchestration skill.** Workers figured it out each time. Patched upstream in the BLS OEWS skill (Round 5) with explicit retry pattern.
7. **FY per diem fallback missing from orchestration skill.** S3 worker hit empty FY27 rates. Patched upstream in Per Diem skill (Round 1) with explicit fallback rule.
8. **CALC+ endpoint sanity check missing.** Workers hit 404s chasing wrong URLs. Patched upstream in CALC+ skill (Round 3).

P2 (opportunistic, quality improvements)

1. **IT PM decision rule.** S1 worker burned a round-trip on 15-1299 (-21% vs CALC+) before pulling 11-3021 (+4%). Patched with decision table: DoD/IC IT PM defaults to 11-3021; civilian dual-pull.

2. **Network Engineer SOC disambiguation.** 15-1241 (architect) vs 15-1244 (sysadmin). Patched with decision rule; default 15-1241 conservative.
3. **SOC-not-at-MSA fallback.** S1 hit 15-1256 not published at Huntsville. Patched upstream in BLS skill (Round 5) with parent-SOC-family rollup.
4. **Productive hours user-override reconciliation.** Skill defaults to 1,880 but users may provide 1,920 in handoff. Patched with explicit "user input wins" rule and back-solve protocol.
5. **Seniority inference for implicit tiers.** S2 worker had to improvise P75 for "Security Project Manager" without explicit seniority label. Patched: cleared/technical PM defaults to P75.
6. **Divergence-triggered SOC re-pick automation + Raw Data retention + Contract vehicle usage rule + Scenario block row formula fix (12→15 rows) + Pre-delivery sanity checklist.** Bundled as Step 8.6 and additions to Step 1 mapping and Information to Collect.

Round 1 patches shipped

All 17 findings above shipped as Round 1 patches in April 2026 immediately following Wave 1 grading. Key additions:

- Rewrote Step 4 to reference CALC+ skill as authoritative endpoint source
- Added DOWNSTREAM CELL REFERENCES map to Step 8
- Added Step 8.5 post-recalc sanity gates (3 checks)
- Added Step 8.6 pre-delivery sanity checklist (14 items)
- Added Materials Handling Fee Decision Gate to Step 5B
- Expanded Sheet 5 Methodology section with full FAR citation set
- Added PM decision table to Step 1
- Added Network Engineer disambiguation to Step 1
- Added Contract Vehicle Usage Rule table tuning burden ranges by vehicle
- Added Productive Hours Reconciliation rule
- Added Seniority inference for implicit tiers
- Added Divergence-triggered SOC re-pick and Raw Data retention requirements
- Fixed Scenario block row formula (12→15 rows)

Round 2 patches queued

None block current ship state. Queued items emerged from grader observation but are not reproducibility bugs:

1. **Workflow A+ SOW/PWS decomposition** not tested in Wave 1. Needs dedicated scenarios.
2. **Workflow B rate validation only** not tested. Needs dedicated scenarios.
3. **24x7 shift coverage (Step 0.5)** not tested. Needs dedicated scenario.
4. **Retest all three Wave 1 scenarios** with Round 1 patches applied to confirm regression-free fix.

Wave 2: Post-MCP migration + Wave 5 FFP inheritance + v2 ai-boundaries gate (Claude Code Desktop, Opus 4.7)

Context

Wave 2 is the first LH/T&M testing round since Wave 1. It consolidated three streams of work into a single ship:

1. **Inheritance of six universal patches derived from FFP Wave 5.** ai-boundaries positioning, pre-flight MCP dependency check, Workflow B data-only rewrite, Step 0 two-stage validation gate, DoD installation to GSA per diem crosswalk, multi-destination travel sheet parameterization, CLI recalc fallback.
2. **v2 ai-boundaries gate.** The original FFP Wave 5 ai-boundaries patch failed a live LH/T&M test. In S3 (described below), the skill 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 ai-boundaries patch. Root cause: the gate lived at Workflow B Step 6 "Stop," which is too far downstream; by that step the model was already in "helpful memo author" momentum. Fix: moved the gate to Step 0 with a token-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).
3. **Two end-to-end scenarios validating workbook production end-to-end** against the post-inheritance skill.

Scenarios

- **S1 FHWA Application Modernization** (DOT civilian IT, 14 FTE, no travel). SOW-driven build (Workflow A+). Exercises SOW decomposition, Step o Stage A/B gate, PM dual-pull decision, civilian-IT wrap preset, 5-year PoP with escalation, CALC+ rate validation on mixed seniority team.
- **S2 Cyber/IR Pentagon** (DoD cleared Secret, 4 analysts, 4 travel destinations: Fayetteville NC, Huntsville AL, NSA Bethesda, San Francisco CA). Structured input build (Workflow A-LH). Exercises DoD installation crosswalk, multi-destination Sheet 5 parameterization, day-trip M&IE (Bethesda), cleared burden preset, FY rollover, small cleared team PM SOC choice.
- **S3 Senior Data Scientist \$225/hr DC rate validation only** (Workflow B). Exercises the ai-boundaries gate. Wave 5 equivalent on FFP triggered the original patch; re-run here against the post-Wave-5 inherited skill is what surfaced the gate-positioning failure described above.

S1 findings (FHWA Application Modernization)

Workbook built cleanly, \$12.5M mid 3-year total, all 3 sanity gates passed (per-FTE in [100K,1M], Sheet 1 total = Sheet 2 mid total, burden multiplier cross-sheet consistent).

PM dual-pull caught divergence: initial SOC 11-3021 landed +34.9% above CALC+ P50 title-match, rejected and re-picked to 13-1082 (Project Management Specialist) which landed -13.1% within the expected tier band. Raw Data sheet retained both SOC queries showing the decision trail.

Evaluator found 10 skill issues, all patched before ship:

1. **Cyber/IR PM SOC rule gap.** Civilian-IT PM rule was clear (dual-pull 11-3021 vs 13-1082). Cyber/IR PM rule was not. Added: cyber/IR PM defaults to 13-1082 (Project Management Specialist) with CALC+ dual-pull for validation.
2. **PM P75 too aggressive for small cleared teams.** Default P75 for PM role produced overpricing when the PM is effectively a lead rather than a layer above a team. Added: P50 default when team size is 6 or fewer, with note to shift to P75 if PM is explicitly a separate management layer.
3. **Step 9 present_files CLI mismatch.** CLI does not have present_files. Worker improvised a file-path report. Codified: Step 9 environment fork (claude.ai / CLI / macOS Numbers), CLI path = absolute file path in response.
4. **Step 8.5 Gate 1 column refs stale.** Gate 1 referenced Sheet 2 columns by letter (D, E, F) after a prior Sheet 2 layout revision shifted burdened-rate column from E to F. Rewrote as named references.

5. **NSA Bethesda crosswalk points to wrong locality.** Was Montgomery County; should be DC composite per GSA convention for NSA Bethesda staff. Fixed in crosswalk table.
6. **FY rollover guidance missing.** Contract PoP starting within 6 months of FY rollover should query both FY rates and note refresh on publication. Added.
7. **Stage A/B gate skip for structured inputs.** Workflow A with SOW/PWS Builder structured handoff does not need Stage A decomposition approval; only Workflow A+ raw SOW text does. Added skip rule.
8. **Secret vs TS/SCI burden split not explicit.** Cleared burden preset was a single row. Split into Secret (2.0-2.2) and TS/SCI (2.2-2.4) rows with note about SCIF overhead not in BLS/CALC+ data.
9. **CALC+ keyword_search returning full corpus for stats-only queries.** Redirect to igce_benchmark for percentile queries where record-level data is not needed.
10. **Tier-matched keyword rule missing.** Query each seniority tier with its own keyword string. Aggregate title-match pools produce false divergence flags when a Senior LCAT is compared against a pool containing Juniors.

S2 findings (Cyber/IR Pentagon)

Workbook built cleanly, \$9.7M mid 3-year total. PM divergence-triggered re-pick caught +53.7% overpricing on 11-3021; re-picked to 13-1082 which landed at +11.7% within the cleared-team premium band. Multi-destination travel sheet built with 4 blocks (Fayetteville, Huntsville, DC composite for NSA Bethesda, SF), day-trip M&IE fired correctly for NSA Bethesda (same-day return from Pentagon). Burden tuning 2.0 / 2.2 / 2.4 landed on the Cleared IDIQ row of the contract vehicle table.

Evaluator found 8 skill issues, all patched:

1. **Cyber/IR PM SOC rule.** Same as S1; both scenarios hit this gap independently. Confirmed the patch.
2. **PM P75 aggressive for small cleared teams.** Same as S1. Confirmed.
3. **Step 9 present_files CLI mismatch.** Same as S1. Confirmed.
4. **Step 8.5 Gate 1 column refs stale.** Same as S1. Confirmed.
5. **NSA Bethesda crosswalk.** Same as S1. Confirmed; S2 would have shipped with wrong lodging rate if the S1 patch had not been in place.
6. **FY rollover guidance.** Same as S1. Confirmed.
7. **Stage A/B gate skip.** Same as S1. Confirmed; S2 used structured input and the Stage A prompt read as unnecessary friction.

8. Secret vs TS/SCI burden split. S2 was Secret; the original single cleared preset row would have nudged burden too high. Confirmed patch.

S1 and S2 corroborated each other on 8 of 10 issues; 2 issues (CALC+ redirect, tier-matched keyword rule) were S1-only but ported across.

S3 findings (rate validation) - ai-boundaries gate failure

S3 exposed the Wave 5 FFP ai-boundaries patch as insufficient in production. The skill, when asked "is \$225/hr reasonable for a Senior Data Scientist in DC," produced:

1. A full Price Reasonableness Determination memo template populated with 5 separate "fair and reasonable" determinations.
2. A recommended negotiation position toward CALC+ P75 ("push back if the vendor can't articulate the clearance value").
3. Draft Evaluation Notice language.
4. An invented 15-25% TS/SCI clearance premium applied as if it were market data.

All four outputs are Tier-1 ai-boundaries violations per the repository's ai-boundaries.md. The Wave 5 patch's instruction "do NOT assert fair/reasonable" sat in Workflow B Step 6; by the time the model reached Step 6 it had already drafted most of the memo via Steps 1-5. The "Stop" instruction read as advisory.

Fix: v2 ai-boundaries gate.

- Moved to Step 0 as a verbatim refusal-template token scan. If the user prompt contains any of `reasonable / fair / defensible / recommend / negotiate / push back / counter / Evaluation Notice / PNM / determination`, the skill emits the verbatim refusal template as its first response and offers two options:
- **Option A:** positioning data only. Skill pulls CALC+ + BLS data, places the proposed rate on the distribution, produces a positioning sheet with neutral labels ("Within CALC+ FFP premium range" / "Metro geographic premium; see Methodology for factor decomposition" / "CO review recommended for factors outside BLS/CALC+ data"). No evaluative verbs in any output.
- **Option B:** memo template fill. Skill produces a memo template with `[CO to complete]` placeholders in the Determination, Conclusion, and Recommendation sections. Skill only fills those sections verbatim if the CO supplies the rationale and conclusion in the prompt.
- Evaluative-verb scrub across all output paths. "Defensible," "reasonable," "acceptable," "competitive," "outlier" removed from Methodology sheet prose, chat summary, validation sheet Status column.

- Out-of-data premiums (TS/SCI, OCONUS hazard, SCIF overhead, specialty labor market) named as gaps; skill flags rather than invents ranges.
- ai-boundaries citation block added at the top of the skill naming the rule explicitly with examples of what the skill does and does not do.

S3 re-run after the v2 gate patch: the skill emitted the refusal template at Step 0, received "Option A" from the test caller, and produced a clean positioning sheet with no evaluative claims. Gate held.

Cross-skill audit: bloat removed

The skill had accumulated redundancies and verbose prose across the inheritance. Audit cut 925 to 832 lines (-93) while shipping all patches:

- **Burden Multiplier Guidance section removed.** Duplicated the Vehicle table in worse form.
- **Step 8.5 and Step 8.6 merged.** 8.5 was "post-recalc sanity gates," 8.6 was "pre-delivery sanity checklist." The 3 gates and the 14-item checklist overlapped on 8 items. Consolidated to a single Step 8.5 with the 3 gates and 6 unique checklist items.
- **Edge Cases reduced to traps list.** Pre-audit Edge Cases mixed genuine silent-wrong-answer traps with quality suggestions. Split: traps stay in Edge Cases, quality suggestions moved to a new "Optional enhancements" appendix.
- **Quick Start Examples cut from 12 to 4.** The 4 retained cover the distinct pricing-structure decision gates. The 8 trimmed were restatements against different agencies.

Wave 2 aggregate

Metric	Value
Rounds	3 (S1, S2, S3)
Workbooks / documents produced	2 workbooks + 1 positioning sheet (post-gate)
Tier-1 ai-boundaries violations identified	1 (S3, pre-v2-gate)
Skill defects identified	18 unique (10 in S1, 8 in S2, 8 overlap)
Skill defects fixed	all 18 + v2 gate
Line delta	925 to 832 (-93)
All 3 sanity gates passed in S1 and S2	yes

What has not been tested

- Wave 1 S1/S2/S3 scenario retest on the Wave-2-patched skill.
- 24x7 shift coverage (Step 0.5) still not exercised.
- Full Workflow B Option A positioning sheet production (covered briefly in S3 post-gate; no extended test).
- Option B memo template fill with CO-supplied rationale.
- Sonnet 4.6 parity.

Queued for Wave 3.

Wave 3 (inherited from CR Wave 1 lazy-prompt testing)

Wave 3 (Cross-skill patches inherited from CR Wave 1 lazy-prompt testing): CR Wave 1 found 22 issues, 14 patched. All 7 universal patches ported to LH/T&M identically to FFP: DOE lab crosswalk rows added, BLS MSA URL fallback, Workflow A ambiguous rule, Step 9 macOS Excel/Numbers branch, BLS wage-cap 10% rule, shift coverage upfront, Methodology depth. Plus 5 editorial fixes: Rate Validation neutral phrasing, Sheet 4 travel include-stub, Stage A/B skip clarification, igce_benchmark default, NAICS/PSC proactive. **Status:** inherited, not re-tested on LH/T&M. LH/T&M remains validated through Wave 2 (S1 FHWA + S2 Pentagon).

Independent grading methodology

Wave 1 testing record produced under consistent methodology:

- Scenarios and assertion matrices committed in writing before any worker output was read
- Grader did not coach workers during runs
- Assertions graded strict on literal wording; soft fails noted explicitly
- Worker self-critiques incorporated as findings when corroborated by observation
- All findings come from direct observation of worker output, not inference from memory

Wave 4 (universal patches inherited from CR Wave 2 detailed-prompt round)

Wave 4 (Universal patches inherited from CR Wave 2 detailed-prompt round): Same 11 universal patches ported to LH/T&M identically to FFP. See Wave 7 on FFP testing record for full list. Status: inherited, not re-tested on LH/T&M directly.

Wave 5 (2 cold tests completed, 2 rate-limited; 2 universal patches shipped)

Wave 5 was a targeted wave driven by two objectives: (1) port horizontal findings from CR Wave 4, and (2) exercise untested territory from Wave 2/3 "What has NOT been tested" list. Four cold tests were launched simultaneously; two completed before Anthropic rate limits cut off the other two.

Tests run

#	Scenario	Workflow	Purpose	Status
1	Lockheed NSA Fort Meade cyber ops, CO-supplied FPRA 2.68	LH	CR Wave 4 horizontal port validation (DCAA/FPRA override rule)	Completed
2	DISA 24x7 SOC + \$680K/yr pass-through materials	T&M	24x7 Step 0.5 math on T&M (untested)	Rate-limited
3	USCIS ELIS modernization from raw SOW	LH (A+)	Workflow A+ SOW decomposition (untested)	Rate-limited
4	Senior Red Team Operator \$285/hr DC "is this reasonable"	Workflow B	ai-boundaries v2 gate + Option A/B (untested on LH/T&M)	Completed

Findings and patches shipped (2)

#	Patch	Section affected	Trigger	Horizontal?
1	CO-supplied DCAA-audited rates override rule (use FPRA verbatim; do NOT bookend ± 0.2 around an audited rate; document source in Methodology)	Contract Vehicle Usage Rule section	Test 1 confirmed the CR Wave 4 gap is universal: "DCAA" / "FPRA" do not appear anywhere in LH/T&M. Generic custom-burden rule handles the input but semantically treats the audited rate as a midpoint, not an authoritative point estimate.	Yes — CR Wave 4 port
2	Workflow B gate fires unconditionally on entry (prior gate was token-gated; a prompt like "validate these rates" bypassed the gate and skipped the Option A/B choice)	Step 0 Workflow B gate	Test 4 surfaced a universal silent-bypass path: Workflow B triggers ("validate these rates," "check this proposal") do NOT contain Step 0 gate tokens ("memo," "determination," "fair and reasonable," etc.), so a worker routes to Workflow B → Step 0 → scan finds nothing → waves through to analysis without ever presenting the Option A/B refusal template.	New (LH/T&M-originated)

Other observations from Test 4 (considered but NOT patched per universal-only discipline)

- **"Reasonable" (standalone) not on memo-drafting token list.** Test 4 worker observed the word "reasonable" alone is not in the token list (only "fair and reasonable," "price reasonableness," "reasonableness memo" are). Substring matching made the gate fire anyway. Addressed by the Patch 2 gate-fires-unconditionally rewrite plus expanding the memo-drafting token list to include "reasonable" (standalone), "validate," "acceptable," "justify."
- **Three duplicated lists of prohibited evaluative verbs** (Operating Principle, Step 0 hard prohibitions, Step 6 stop) could be consolidated. Skipped: editorial cleanup, not a correctness gap; three reinforcing lists are a feature, not a bug.

- **MCP output field outlier_bounds_2sigma passes through unfiltered.** MCP-side naming, not skill narrative. Too narrow to warrant skill-level guidance.

Findings from Test 1 bundled into the DCAA override patch

- **Low/High bookending semantics when CO rate is authoritative.** The ± 0.2 bookending produces fictional scenario display when the CO supplies an audited rate (DCAA FPRA 2.68 IS the rate, not a midpoint). Patch 1 explicitly states: apply as point estimate (single column); if a band is still shown, label as "sensitivity display only; FPRA is authoritative."
- **Methodology source-citation for overridden defaults.** Patch 1 requires documenting FPRA effective date, approving authority, rate composition.
- **Divergence between CO-supplied rate and vehicle-table band.** Patch 1 states: trust the CO-supplied rate and note the divergence in Methodology rather than reconciling to the table.

Wave 5 completion (2 rate-limited tests retried once rate limits cleared)

After the initial 2 cold tests, Tests 2 (24x7 T&M shift coverage at DISA Fort Meade) and 3 (Workflow A+ SOW decomposition on USCIS ELIS modernization) were rate-limited by Anthropic. Retried approximately 2 hours later. Both completed cleanly. Summary of all 4 Wave 5 cold tests:

#	Scenario	Result on Wave 5 patches	New universal gaps surfaced
1	Lockheed NSA Fort Meade CPFF 2.68 FPRA	DCAA override patch validated; 2.68 used as point estimate with ± 0.2 bookending labeled "sensitivity display only"	None
2	DISA 24x7 SOC + \$680K pass-through materials	Both patches correctly abstained (Workflow A, no FPRA); 4.2 FTE single-seat 24x7 math cascaded cleanly through T&M block layout; materials at cost separated from burdened labor	Gate 3 cell reference bug (see Patch 3 below)
3	USCIS ELIS modernization, raw SOW Workflow A+	Both patches correctly abstained; Stage A/B gate held; 7 LCATs / 15 FTE decomposition defensible; \$13.9M 3-yr mid total	Offset convention ambiguity (same as Test 2, different manifestation)
4	Senior Red Team \$285/hr DC Workflow B	Workflow B gate fired unconditionally on "reasonable"; Option A produced cleanly with no evaluative verbs	None

Patch 3: Gate 3 cell reference correctness fix (Wave 5 completion)

Tests 2 and 3 both surfaced confusion around the Sheet 2 block-offset convention. Test 2 went further and identified a concrete silent-wrong-answer bug: the Sheet 2 block layout (line 648-658) and the Step 8.5 Gate 3 cell reference (line 781) were inconsistent.

- **Block layout** said +7: Burdened Mid (row 7 for block 1 under the natural "+N = N-th row of block" convention).
- **Gate 3** said row +9 = Burdened Mid at B9.

Under either possible interpretation of "+N", Gate 3 was reading the wrong cell: B9 is either a blank separator row (under "+N = Nth row") or Burdened High (under "+N = block_start + N"). Any worker running Gate 3 verbatim was either getting a false-fail (sheet 1 burden of 2.2 vs sheet 2 blank == divergence) or, worse, silently comparing Sheet 1 Mid to Sheet 2 High.

Fix: 1. Block layout now shows explicit row numbers for block 1 in parentheses next to each offset (row 1 header, row 7 Burdened Mid, etc.). The offset convention is called out explicitly: +K is the K-th row of the block where +1 is block_start itself. 2. Gate 3 cell reference corrected from B9 to B7, with

a note calling out that prior versions referenced the blank separator row as a silent-wrong-answer bug. 3. Per-block formula added: for block N, the Burdened Mid cell is $B\{7 + (N-1)*15\}$.

Verified inert on FFP and CR: FFP doesn't have an equivalent Gate 3 cell-compare pattern; CR's block layout uses a different convention ($+K = \text{block_start} + K$) that is internally consistent with its concrete row examples (rows 2-19 for block 1 with $+1 = \text{BLS Base at row 2}$). No port needed.

Wave 5 universal-only discipline: other observations skipped

Tests 2 and 3 also surfaced:

- **Period row offsets assume 4 periods (+10 to +13).** When PoP is 3 years, +13 is blank. Cosmetic. Skip.
- **Scenario Range summary row placement unspecified** (Sheet 1 vs Sheet 2, top vs bottom). Worker improvised successfully in both tests. Consistency issue, not correctness. Skip.
- **FTE annotation row +9 undefined** (would overlap with blank separator under the corrected convention). Cosmetic. Skip.
- **Cross-LCAT Low/High aggregation cells (I2/I3) undocumented.** Worker improvised. Consistency issue, not correctness. Skip.
- **Materials vs ODC placeholder band co-location.** T&M has real materials and placeholder ODCs on the same sheet; skill doesn't prescribe their visual separation. Minor. Skip.

All five were considered and deliberately dropped per the universal-only + "patch correctness bugs, not documentation polish" discipline. The Gate 3 bug was the only one that rose to the bar.

Wave 5 line delta

SKILL.md: 832 → 859 lines (+27 across all 3 Wave 5 patches: DCAA override +~10, Workflow B gate unconditional fire +~5, Gate 3 correctness fix +~12). Ceiling remains 1,000.

Wave 5 regression summary

- **4 of 4 cold tests completed** (2 initial + 2 retried after rate-limit reset)
- **3 universal patches shipped:** DCAA/FPRA override rule, Workflow B gate unconditional fire, Gate 3 cell reference correctness fix
- **Zero patch misfires observed** across all 4 scenarios
- **Zero new universal gaps** beyond the one addressed by Patch 3

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