INDEPENDENT EVALUATION

BLS OEWS API Testing Record

Federal acquisition skill validated across sixteen runs in two testing waves on two Claude models; Rounds 2 through 4 patches shipped

Skill
bls-oews-api (post-Round 4 patches; 112/112 assertions passed + 8 live-verified patches)
github.com/1102tools/federal-contracting-skills

Date
April 2026
1102tools.com

Part 1: For Federal Acquisition Users

The bottom line

Two waves of independent testing in April 2026 (16 end-to-end runs, 224 binary assertions graded) show the BLS OEWS API skill reliably pulls market wage data across four real-world federal acquisition scenarios on both Claude Opus 4.7 and Claude Sonnet 4.6. One assertion failed in Wave 1 (Sonnet using a stale Knoxville MSA code that silently returned empty data, causing an unnecessary fallback to national and a 27% underreported wage). The skill was patched. Wave 2 retested the same four scenarios on both models and all 112 assertions passed.

Wave 2 aggregate: 100% pass across 112 assertions.

Scenarios tested and how reliably they work

Scenario	Models	Result
Full wage distribution, standard metro and occupation (DC Information Security Analyst)	Opus, Sonnet	Reliable both waves
Metro fallback to state, off-common-table occupation, wage cap (Oak Ridge Nuclear Engineer)	Opus, Sonnet	Wave 1: Sonnet failed on wrong MSA code. Wave 2: Reliable both models after patch
Multi-metro comparison, federal-installation markets (Baltimore + Colorado Springs vs. national cyber)	Opus, Sonnet	Reliable both waves
High-wage occupation hitting the $239,200 reporting cap (SF Software Developer P90)	Opus, Sonnet	Reliable both waves

Manual-verification checklist

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

1. MSA code is correct for the metro you meant. When BLS returns "series does not exist" for all datatypes at your metro, the value could be a real non-publication OR a wrong area code you typed. The Wave 2 patch added Critical Rule #10 to the skill: verify the MSA code against the current BLS list before assuming non-publication. Cross-check by trying a common SOC (15-1232 Help Desk) at the same metro. If THAT also returns empty, you probably typed the wrong code.

2. Cap vs. suppression are different and the skill distinguishes them. When a percentile returns –: - With footnote text containing "115.00" or "239,200": CAP. The true wage is at least $239,200. Use as lower bound. Document in your IGCE narrative. - With footnote "Estimate not released" or RSE text: SUPPRESSION. The value is unknown. Fall back to broader geography or related SOC.

3. BLS OEWS does NOT publish hourly percentiles as discrete series. The skill documents datatype codes 06/07/09/10 for hourly P10/P25/P75/P90. These work, but BLS computes them as annual/2080. For contractor billable-hour pricing, divide annual wages by 1,880 instead.

4. National vs. state vs. metro fallback order is NOT always "broader is safer." The skill added explicit guidance in Wave 2: when an occupation is concentrated in a few dominant employers in a state (nuclear engineering in TN, certain intel SOCs in VA), BLS may suppress at the state level while publishing clean metro data. Don't fall through to state if metro worked.

5. Burden multipliers are indicative, not certified. The skill offers 1.5x to 3.0x bands for IGCE rate derivation. These are practitioner consensus ranges, not DCAA-certified figures. For defensibility in contested price negotiations, cite an actual vendor-specific wrap rate or published DCAA guidance, not just the skill default.

Choosing between Opus 4.7 and Sonnet 4.6

Short answer: both work. Use either. Differences are small.

Opus 4.7 hits BLS API more efficiently per tool call, recovers from parameter errors faster when they occur, and in testing caught a capped P90 that a Wave 1 Sonnet run missed (because Sonnet only queried up through P75 by default). Opus also tends to add supplementary context like sanity checks against related metros, which is useful for IGCE narrative but strictly optional.

Sonnet 4.6 is faster on simple queries and produces cleaner output in fewer tokens. The Wave 1 silent-failure bug (wrong Knoxville code) was a Sonnet-only issue that is now fixed in the skill. On the patched skill, Sonnet and Opus produce comparable quality output.

Neither model hits tool-use limits on BLS queries. This skill is light; both models complete in a single response.

What the skill does not do

- **It does not benchmark individual employer wages.** OEWS aggregates employer-reported data; it does not provide data on specific companies.
- **It does not replace GSA CALC+ for price reasonableness.** BLS reports base wages paid to employees; CALC+ reports awarded ceiling rates including fringe, overhead, and profit. Use both.
- **It does not handle classified or national-security occupations with special OPM paybands.** OEWS is private-sector focused; for cleared federal employee positions, supplement with OPM locality pay tables.
- **It has not been tested on** hourly P10 (datatype 06) at a non-standard metro, industry-specific queries at the state level (the skill correctly flags industry queries as national-only), or year-over-year trend analysis across the 2024 MSA boundary realignment (where same codes may cover different counties).

Environmental gotchas on claude.ai web chat

Gotcha	What happens	Workaround
Wrong MSA code silently returns "series does not exist"	Model may fall back to state or national without flagging the code error	Worker (and Rule #10) now prescribes verification. If output cites national when you asked for a metro, check the MSA code first.
BLS v1 API has 25 queries/day limit	Unkeyed runs exhaust fast	Register for a free v2 key (500 queries/day, 50 series per query = effectively unlimited) at https://data.bls.gov/registrationEngine/
OEWS publishes annually in April	Running close to release date may miss new data	Skill's `detect_oews_year()` helper probes for newer data. Skip for single-shot queries before April 15; call for multi-query sessions after

Part 2: For Developers and Technical Reviewers

Testing methodology

Scenarios

Four scenarios were selected before any testing began, chosen to exercise distinct capabilities and common failure modes:

- **S1 — Full wage distribution:** DC Information Security Analyst (SOC 15-1212). Common-metro query, well-trodden SOC, full percentile and hourly distribution. Exercises datatype coverage, common-metro lookup, happy path.
- **S2 — Metro fallback and wage cap:** Oak Ridge Nuclear Engineer (SOC 17-2161). Off-common-table SOC, small metro at risk of suppression, specialty market with top-coded percentiles. Exercises fallback logic, cap detection, silent-failure detection on wrong MSA codes.
- **S3 — Multi-metro comparison:** Information Security Analysts across Baltimore-Columbia-Towson MSA, Colorado Springs MSA, and national. Exercises multi-query orchestration, 2020-census MSA codes, comparative wage variance presentation.
- **S4 — Wage cap edge case:** San Francisco Software Developer P90 (SOC 15-1252). Exercises cap handling, footnote code 5 distinction from suppression, interpretation of capped values for IGCE defensibility.

Each scenario had a 14-point binary assertion matrix: 8 general (G1-G8, same across all scenarios) and 6 scenario-specific. Assertions were written before any worker output was seen; assertions were not revised after the fact based on what workers produced.

Environment

- claude.ai web chat, fresh conversation per run
- Skills installed: `bls-oews-api` and `bls-oews-api-reference` in Wave 1; `bls-oews-api` (merged) in Wave 2
- Models: one run on Opus 4.7, one run on Sonnet 4.6, per scenario
- Total: 4 scenarios × 2 models × 2 waves = 16 runs

Grading

The grader (Claude Code session separate from any worker run) read only the worker's final response text. Workers were not coached during runs. Each assertion graded binary pass/fail. Suspicious details were noted even when the assertion passed.

Wave 1 results (pre-patch)

Scenario	Sonnet 4.6	Opus 4.7
S1 DC InfoSec full distribution	14/14	14/14
S2 Oak Ridge Nuclear + fallback	**13/14**	14/14
S3 Multi-metro comparison	14/14	14/14
S4 SF Software Dev P90 cap	14/14	14/14
Total	**55/56 (98%)**	**56/56 (100%)**

Wave 1 aggregate: 111/112 (99%).

Single failure: S2.2 Sonnet — Wrong Knoxville MSA code

Sonnet queried Knoxville with area code `0028700`. The API returned "series does not exist" for every datatype. Sonnet interpreted this as legitimate non-publication and fell back to Tennessee state (which was genuinely suppressed) and then to national ($127,520 median).

The correct Knoxville MSA code is `0028940`. Opus used it in Wave 1 and returned the actual Knoxville-specific median of $174,380. The Sonnet answer was understated by 27%.

The BLS API returns the same "series does not exist" response for both (a) a typo in the area code and (b) a combination that BLS genuinely did not publish. Worker had no signal to distinguish the two.

Wave 1 findings: 7 cross-run patterns worth patching

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

1. **Hourly percentile codes existed but were undocumented.** Sonnet S1 claimed hourly percentiles don't exist and derived them via annual/2080. Opus S1 inferred codes 06/07/09/10 from the pattern, queried them, got data. Values match annual/2080 exactly, but the discrete series codes exist.
2. **Quick Start default omitted P25/P75.** Workers consistently needed P25/P75 for interquartile range reasoning. Default pulled only mean/median/P10/P90.
3. **Common metros table was federal-underweighted.** Defense hubs (Colorado Springs, Huntsville, Norfolk, San Antonio) and DOE labs (Knoxville, LANL, Richland, Idaho Falls) were absent. Workers compensated via memory or external lookup.
4. **Cap vs. suppression conflated.** Skill's `format_oes_value()` returned `[Capped]` for any `–` value regardless of footnote text. Both cases look the same to a naive parser but have different IGCE-defensibility implications.
5. **No composite fallback helper.** Every caller reinvented metro → state → national logic.
6. **MSA-code verification step was missing.** The Knoxville 28700 silent-failure bug had no prescriptive counter in the skill.
7. **Engineering SOC block missing from common occupations.** Nuclear Engineer (17-2161), Aerospace (17-2011), Mechanical (17-2141), etc. were absent, and the default "Systems Engineer → 15-1211 Computer Systems Analysts" mapping is wrong for non-IT engineering requirements.

Patches shipped before Wave 2

All patches applied to `~/.claude/skills/bls-oews-api/SKILL.md` before Wave 2 runs began. The `bls-oews-api-reference` split-skill architecture was merged into the main SKILL.md in the same pass.

1. Hourly percentile codes 06/07/09/10 documented in the main datatype table with annual/hourly field explicitly distinguished.
2. Quick Start expanded to pull all 7 IGCE-relevant measures (employment + annual mean + all 5 annual percentiles).
3. `quick_wage_lookup` area parameter changed from optional (`area="0000000"`) to required (no silent national default).
4. Common metros expanded with federal subsections: Federal Capital Region / East Coast, Defense Hubs, DOE Labs, Southeast. Twelve net-new metro codes including Knoxville 0028940, Colorado Springs 0017820, Huntsville 0026620, Santa Fe (LANL) 0042140, Richland WA 0028420, Idaho Falls 0026820.

5. New `interpret_value()` helper distinguishes CAP (footnote code 5) from SUPPRESSION (Estimate not released) and returns distinctly labeled strings.
6. Dedicated "Interpreting Capped Wages" section with practical IGCE guidance: use P75 as uncapped ceiling when P90 is capped, cross-reference CALC+, document the cap in narrative.
7. New `query_with_fallback()` composite helper: metro → state → national with explicit logging of which geography returned the final value.
8. "Fallback Pattern" section explaining the counterintuitive case where state is suppressed but metro publishes (single-employer-dominated occupations).
9. Engineering SOC block added: 172011 Aerospace, 172031 Biomedical, 172041 Chemical, 172051 Civil, 172071 Electrical, 172072 Electronics, 172081 Environmental, 172112 Industrial, 172141 Mechanical, 172161 Nuclear, 172199 Engineers (All Other) for non-IT Systems Engineer roles, 172171 Petroleum.
10. Critical Rule #10 added: verify MSA code before assuming suppression. If a metro query returns "series does not exist" for EVERY datatype, check the area code against the current BLS MSA list before falling back.
11. 2024 MSA realignment note (OMB Bulletin 23-01) added to the metros section: same code may cover different counties vs. pre-2024 data.
12. `format_oes_value` updated to distinguish CAP vs SUPPRESSED in returned strings.
13. Full state FIPS appendix expanded from ~10 states to all 52 (50 states + DC + PR).

Wave 2 results (post-patch)

Scenario	Sonnet 4.6	Opus 4.7
S1 DC InfoSec full distribution	14/14	14/14
S2 Oak Ridge Nuclear + fallback	**14/14**	14/14
S3 Multi-metro comparison	14/14	14/14
S4 SF Software Dev P90 cap	14/14	14/14
Total	**56/56 (100%)**	**56/56 (100%)**

Wave 2 aggregate: 112/112 (100%). Wave 1 silent-failure bug fixed.

Methodology upgrades observed beyond the matrix

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

- **Sonnet S1:** used hourly percentile codes 07/09/10 directly instead of deriving via /2080
- **Sonnet S2:** used correct Knoxville MSA 0028940 from expanded common metros table; documented cap on P75 (new) alongside cap on P90
- **Sonnet S3:** used Colorado Springs 0017820 from table (Wave 1 inferred from OMB pattern)
- **Sonnet S4:** used the "scale maxing out at 500 lbs" analogy for cap interpretation — more useful than Wave 1's textbook explanation
- **Opus S2:** caught P90 cap alongside P75 cap (Wave 1 only flagged P75); noted the counterintuitive "state suppressed, metro published" pattern explicitly
- **Opus S3:** added employment RSE (2.0%) and wage RSE (1.4%) for data quality reporting
- **Opus S4:** extended cap interpretation with practical decision tree (use P75 as upper anchor, cross-reference CALC+, hourly fallback)

Truncation investigation

Three Wave 2 workers (Sonnet S3, Opus S1, Opus S2) reported that lines 185-563 of the 746-line merged SKILL.md were "cut off in my view" — implying web-chat was truncating half the file. One Wave 2 worker (Opus S4) referenced content from line 689+ explicitly.

A targeted audit on a separate Opus Wave 2 session asked the worker to quote verbatim from three specific sections in the claimed cut zone: 1. Idaho Falls area code (line ~212) 2. Critical Rule #10 (line ~310) 3. `query_with_fallback()` function signature (line ~395)

Worker returned all three verbatim with correct locations. **Truncation is not real web-chat behavior.** Workers who claimed truncation were describing "I didn't re-read the full file to check this specific content" in imprecise language. The 746-line file loads fully.

This matters for downstream skill design: merged skill files up to at least 750 lines are safe to ship without architectural concern for claude.ai web chat delivery.

What was not tested

- Hourly P10 (datatype 06) at any metro
- State-level queries with non-000000 industry (skill correctly rejects these as national-only, but this behavior was not explicitly exercised)

- Year-over-year trend queries across the 2024 MSA boundary realignment
- Occupations at the $239,200 annual cap that DO publish a numeric value (vs. the SF Software Dev P90 case which was a clean cap, and the Oak Ridge Nuclear P75 cap — both were – with footnote 5)
- The `detect_oews_year()` probe function (Opus worker flagged that it may query an invalid series ID `OEUN00000000000000000004`; not verified against live API)
- Parsing of RSE datatypes 02 and 05 (not in Quick Start, not exercised directly)
- Series that return ∗ (RSE > 50%, unreliable) — skill documents the behavior but no test scenario triggered it
- Series that return # (withheld) — skill documents the behavior but no test scenario triggered it

Round 2 patches shipped

These emerged from Wave 2 self-assessments. Shipped before publication and validated via a Wave 3 sanity check (single Opus run against Idaho Falls Nuclear Engineer query exercising all Round 2 patches; no regressions).

1. **Footnote code-match as additional check alongside text-match** in `interpret_value()` and `format_oes_value()`. Durable against future BLS text reformatting; text match preserved as fallback.
2. **RSE datatypes 02 and 05 added to Quick Start.** Default now pulls 9 measures including employment RSE and wage RSE.
3. **"When P90 is capped" decision tree** added as subsection of Interpreting Capped Wages. Five prescriptive steps: P75 as uncapped ceiling, cross-reference CALC+, national P75/median ratio for derived estimate, commercial surveys for tech markets, hourly conversion for sanity check.
4. **Common metros reordered to lead with federal installations.** Order: Federal Capital Region / East Coast → Defense and Intelligence Hubs → DOE National Lab Metros → Major Commercial → Southeast. Added Warner Robins (49660) to defense hubs; Albuquerque/Sandia (10740) to DOE labs.
5. **`bls_to_igce_rate()` parameterized with `productive_hours`** defaulting to 2080 (BLS standard) with explicit 1880 alternative for contractor billable-hour basis. Docstring and example show both.
6. **`detect_oews_year()` verification deferred.** Probe series ID may be invalid; not verified against live API. Held for Round 3 pending actual API test.

7. **Capped result worked example added** to Query Recipes showing the API response shape, `interpret_value()` output, and IGCE narrative template.
8. **CSA vs MSA note added** with Knoxville CSA 34100 vs MSA 28940 example. BLS OEWS uses CBSA codes.
9. **Single-employer-dominated occupation heuristic added** to Fallback Pattern with five known cases (Nuclear in TN, Nuclear/physicists in NM/ID/WA, cleared intel in VA/MD, aerospace in AL Madison, astronomers in MD).

Bonus patches shipped alongside: - Burden multiplier disclaimer in `bls_to_igce_rate` docstring ("practitioner consensus, not DCAA-certified; cite vendor-specific wrap rate or published DCAA guidance for defensibility"). - `format_oes_value()` updated to handle RSE codes 02/05 (returns percentage format).

Wave 3 sanity check

Single Opus run validated Round 2 patches against a query specifically designed to exercise 5 patches at once:

Query: "Pull BLS OEWS wage distribution for Nuclear Engineers in Idaho Falls, ID. Include RSE for employment and wages. If P90 is capped, walk through the decision tree. Compute burdened rate at 2.0x using 1880 contractor hours."

Validation outcomes: - Idaho Falls 0026820 used directly (patch 4: metro reorder confirmed) - Wage RSE 1.0% and employment RSE both queried and reported (patch 2: RSE in Quick Start confirmed) - Cap decision tree explicitly checked and noted "not triggered" with reasoning (patch 3 confirmed) - `bls_to_igce_rate` used with productive_hours=1880 for all burdened calculations (patch 5 confirmed) - Single-employer suppression predicted correctly: "INL dominates the metro sample, so BLS suppresses the headcount" (patch 9 confirmed) - Burden multiplier disclaimer cited in output (bonus confirmed)

New finding from Wave 3 (Round 3 candidate): Footnote code 8 ("Estimate not released" on employment count specifically) is not in the skill's Critical Rules table. Distinct from code 8-as-RSE>50% and code 5-as-cap. Shows up in real data. Queue for Round 3.

Round 3 patches shipped

Shipped based on findings from IGCE FFP Wave 2 testing (April 2026) where an Opus worker burned significant time brute-force-scanning the BLS API to discover Cleveland's MSA code had

been renumbered.

1. **Cleveland MSA renumbered from 17460 to 17410.** Added Cleveland 0017410 to common metros under a new "NASA research centers" subsection with explicit annotation: "renumbered from 17460 in May 2024 OEWS per OMB Bulletin 23-01."
2. **Expanded 2024 MSA realignment note** to document three distinct effects of OMB Bulletin 23-01: same-code composition changes, complete renumberings, and renamings. Added silent-failure pattern warning: if a previously-working MSA code returns NO_DATA across every SOC, the metro was renumbered. Do NOT assume BLS suppressed the occupation. Added BLS area code list URL as the canonical source.

Round 4 patches shipped

All 8 items from Wave 2 Opus self-assessments now shipped. Two required live API verification (items 7 and 8), which was performed against the BLS public API in April 2026.

1. **Footnote code 8 documented as distinct failure pattern.** Table now distinguishes four cases: CAP (code 5), SUPPRESSED (code 8, covers both confidentiality and RSE>50%), RSE flag (∗ value), and NOT PUBLISHED. Skill documents that code 8 text cannot always distinguish confidentiality vs sample-size suppression; treat both as unknown.
2. **Cap variability within single-employer patterns** documented with Knoxville (P75 caps) vs Idaho Falls (P90 publishes clean) contrast. Skill now instructs workers to query before assuming the cap pattern.
3. **`quick_full_distribution` sibling function added** to Quick Start. Pulls all 15 datatypes (annual + hourly percentiles + RSEs) in one call for LH/T&M rate validation or dual-view analysis.
4. **Burden multiplier by contract vehicle table added** to IGCE Rate Derivation section. Nine rows: GSA MAS commercial/cleared, Agency BPA non-cleared/cleared, DoE M&O/FFRDC, DoD prime non-cleared/SCIF, OCONUS, R&D CR. Explicit disclaimer that these are anchors not certified rates.
5. **1880 hours promoted to default** in `bls_to_igce_rate` function signature. 2080 moved to opt-in for direct-labor-rate basis only. Example output updated to show 1880 result first.
6. **RSE interpretation rubric added** as new section. <5% cite as-is, 5-15% cite with range, >15% directional only, ~50% unreliable. Explicit instruction to pull RSE datatypes alongside wage measures for audit trail.
7. **`detect_oews_year()` probe series verified against live BLS API** (April 2026). The 25-character `OEUN00000000000000000004` returns $67,920 for 2024 national all-occupations

annual mean. Docstring updated with verification note.

8. **Dayton MSA renumbering verified and shipped.** Dayton moved from 19380 to 19430 and was renamed Dayton-Kettering-Beavercreek per OMB Bulletin 23-01. Previous 19380 query returns "Series does not exist"; new 19430 returns $107,720 InfoSec Analyst median. Common metros table updated. 2024 MSA realignment note consolidated: both Cleveland (17460→17410) and Dayton (19380→19430) confirmed via live API.

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 for each run is auditable in the `igce-ffp-wave1-runbook.md` and `bls-oews-wave1-runbook.md` source files
- All findings come from direct observation of worker output, not inference from memory of prior sessions

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