

Screening New York State Certified Payrolls Against Published Prevailing-Wage Schedules

Method, Validation, and a Prioritized Candidate Set from Public Data

Concourse Technologies — Research
thomas@concoursetech.com

July 21, 2026

Abstract

New York State publishes two datasets that, taken together, permit a systematic compliance screen that today is performed only by complaint: the certified-payroll wage lines filed with the Department of Labor’s Bureau of Public Work (published on `data.ny.gov`), and the Department’s own prevailing-wage rate schedules. We joined 588,505 public-work wage lines (January–July 2026) against a rate library built from the effective-dated schedules, using conservative rules at every step. Of 554,606 screenable journeyworker lines, 24,947 (4.5%) report a cash wage below the lowest scheduled journeyworker rate mappable to their trade, county, and week. A second screening pass built entirely from the same public sources explains away much of that signal — 26.0% of flagged lines exactly match published *apprentice* rates, and 5.5% exactly match a different classification’s journeyworker rate — leaving a standing set of **6,835 high-confidence findings across 468 contractors, approximately \$1.75M in indicative cash-wage shortfall**. The screen independently flagged the one firm in the data with an adjudicated willful-violation record (EGL Drainage Solutions & Repairs, debarred July 2025), with every cited rate hand-verified against the schedule PDF. All findings are allegations that prioritize human review; none is a determination. An appendix describes how the same analysis is performed for New York City, where a completed pilot demonstrates the method on the Comptroller’s \$220 certified-payroll database.

1 Background

Article 8 of the New York Labor Law requires contractors on public work to pay workers no less than the prevailing rate of wage and supplements determined for their trade and locality. Enforcement is largely complaint-driven: a worker or union files, an investigator requests certified payrolls, and a case develops over months or years. Yet the raw materials of a systematic screen are already public. The Bureau of Public Work publishes certified-payroll wage lines as an open dataset, and the Department of Labor publishes the rate schedules those payrolls must satisfy.

This project asks a narrow question: *using only data the State already publishes, how far can a defensible screen get?* The answer is a two-tier pipeline that ranks every below-floor wage line, removes the classes of false positives we could prove exist, and produces a prioritized, citation-backed candidate set for human investigation — with every screening decision logged and reproducible.

Two framing rules govern everything that follows.

- **Findings are allegations, not determinations.** A certified-payroll line below a scheduled floor is consistent with underpayment, but also with legitimate explanations invisible to this data (most importantly apprentice registration and supplement arrangements). Every output of this screen is a screening candidate requiring investigation.
- **Conservative at every join.** Where a mapping is ambiguous, the screen uses the interpretation *most favorable to the contractor*: the lowest plausible scheduled rate, the earliest rate column of the covering schedule year, and exclusion of any line whose trade or county cannot be resolved with evidence. Unresolved lines are logged with a reason, never silently dropped.

2 Data

2.1 Certified-payroll wage lines

The screen uses the certified-payroll dataset published by the Bureau of Public Work on [data.ny.gov](https://data.ny.gov/dataset/w2zp-sf2x) (dataset `w2zp-sf2x`). We harvested 588,505 wage-bearing worker-week lines with week-ending dates from January 1 through July 20, 2026, across 3,503 contractor accounts, with exact reconciliation against the source (zero dropped rows). Each line reports contractor, project, work classification, county-less project location, straight-time and overtime hours and hourly rates, and weekly gross wages.

The dataset has four documented limitations that shape the entire method:

1. **No supplements field.** The published lines carry cash wage only. Article 8 requires wage *and* supplements; this screen addresses the cash-wage floor only, and says nothing about benefits.
2. **The `job_title` field is unreliable.** It takes two values, and apprentices demonstrably appear as “Journeyman”: our pilot found lines whose pay exactly equals published first-term apprentice rates filed under the journeyworker label. Section 3.3 addresses this directly.
3. **No county field.** The schedules are county-resolved; the payroll lines are not. County must be derived (Section 3.1).
4. **The wages field is the worker’s full weekly gross across all work**, while the hours fields are project-specific. Verified empirically: only 3.6% of all lines satisfy $\text{wages} \approx \text{rate} \times \text{hours}$, with clear week-splitting in the remainder. A reconciliation mismatch therefore carries *no* signal, and the screen never uses one against a contractor.

2.2 Rate schedules

The Department publishes prevailing-wage schedules per county and schedule year (July–June). Two vintages cover the data window: the 2025–2026 determination (covering work through June 30, 2026 — 93% of the lines) and the 2026–2027 determination. The current year’s schedules are served per county by the Department’s wage-schedule application; the prior year’s determination is preserved as the Department’s archived statewide Article 8 schedule. Both are born-digital PDFs parsed with a layout-preserving text extraction, yielding a rate library of 22,128 effective-dated journeyworker rows (11,157 for 2025–2026; 10,971 for 2026–2027) across all 62 counties, and — for the second-tier screen — 26,033 apprentice-rate rows parsed from the schedules’ registered-apprentice sections.

Vintage discipline. Every wage line is screened against the schedule determination *in force on its week-ending date*, and within that determination, against the earliest (lowest) rate column — so a mid-year step increase can never manufacture a shortfall. The New York City pilot (Appendix A) quantifies why this matters: comparing early-2026 work against current-year schedules generated 74 false underpayment signals on an otherwise fully compliant corpus.

3 Method

3.1 County attribution

Project ZIP codes resolve 96.5% of lines to a county via the Census ZCTA–county correspondence. The remainder are attributed by a strict fallback ladder — ZIP-3 prefixes whose New York ZCTAs all lie in one county; project city names that match exactly one county subdivision in the Census gazetteer; city names with overwhelming ($\geq 95\%$) agreement among already-attributed lines; and public-work jurisdictions that name their county — each line carrying a provenance tag for which rule attributed it. A guard disables any gazetteer answer that attributed data contradicts (same-named villages are invisible to the gazetteer; 31 names were disabled). Final coverage: 99.8%. The 943 unresolved lines are genuine multi-county placeholders (“Various Locations”) and are logged, not screened.

3.2 Tier 1: the floor screen

The Socrata work classification is coarser than the schedules’ classifications, so the screen maps each line’s trade to a *family* of schedule classifications and takes the **minimum** journeyworker rate in the family as the floor — under-flagging by construction. Where a reported category has no same-named schedule block, a small curated alias layer applies, each alias cited to the Department’s own classification taxonomy and to schedule scope text (for example, upstate “Steamfitter” work is covered under combined Plumbers & Steamfitters schedules in the 55 counties that publish no standalone Steamfitter block; aliased floors carry a provenance tag naming the taxonomy entry). A line is flagged when its straight-time cash rate falls below the covering floor:

$$\text{flag if } \text{rate}_{\text{paid}} < \min_{c \in \text{family}} \text{rate}_{c, \text{county}, \text{vintage}}^{\text{scheduled}}$$

Of 554,606 screenable journeyworker lines (cash rate \$12–\$150), **24,947 lines (4.5%) across 928 contractors** fall below their floor; 12,602 are high-confidence (shortfall $\geq 15\%$ of the floor). All 1,709 unscreenable lines are logged with reasons (943 unresolvable county; 483 building-service refuse categories outside Article 8; 91 “Welder — as per trade,” whose rate is not determinable from this source; 192 survey-crew lines in five counties whose schedules publish no survey block).

3.3 Tier 2: removing what public data can explain

The pilot proved two false-positive classes exist, so the second pass screens for them statewide — annotating, never deleting.

Apprentice-rate match (downgrade). The schedules publish apprentice terms — percentages of the journeyworker rate, or explicit dollar term rates. A flagged line whose pay

matches a scheduled apprentice rate of its own trade family ($\pm\$0.02$) is most plausibly a registered apprentice filed under the dataset’s unreliable journeyworker label. **6,486 flagged lines (26.0%) across 354 contractors** match; they drop to the bottom of review priority. Hand-verified examples: a line paid \$22.25 equals the Erie County electrician first hour-term exactly; a line paid \$28.98 equals the Suffolk County heavy/highway laborer first term (60% of the \$48.29 base wage printed in the schedule). A match is consistent-with, not proof-of, registration — registration records are not public.

Alternate-classification match (caution). A flagged line whose pay exactly equals ($\pm\$0.01$) a *different* classification’s journeyworker rate in the same county and vintage was possibly filed under a miskeyed category: 1,379 lines (5.5%).

Weekly-gross corroboration (one-sided). Where the weekly gross *does* reconcile with rate $\times$ project hours (ratio 0.98–1.02), the low rate is not a data-entry artifact: 370 lines carry this corroboration. Per the semantics finding above, non-reconciliation is never used as a negative signal.

The standing candidate set — high-confidence findings with no downgrade or caution — is **6,835 findings across 468 contractors, approximately \$1.75M** in indicative cash-wage shortfall (shortfall $\times$ project straight-time hours). Exposure is concentrated: the top 20 contractors account for 45.8%.

Stage	Lines	Contractors
Wage-bearing lines harvested	588,505	3,503
Screenable journeyworker lines	554,606	—
Below scheduled floor (Tier 1)	24,947	928
of which high-confidence ($\geq 15\%$ short)	12,602	605
less apprentice-rate matches	–6,486	—
less alternate-classification matches	–1,379	—
Standing high-confidence candidates	6,835	468

Table 1: The screening funnel. Tier-2 subtractions shown against all flagged lines; the standing row applies both filters within the high-confidence subset.

4 Validation

4.1 An adjudicated corroboration — honestly bounded

The Department’s public debarment list (July 2026 edition; 164 entities) is the only public case-level ground truth. It is structurally the *wrong instrument* for measuring an active-payroll screen — debarment removes firms from public work, so debarred firms exit the certified-payroll population — and an earlier loose name-matching exercise that suggested “56% recall” was discarded once inspection showed the matches rode on generic tokens (*Construction, Group*); we treat that episode as a required hygiene lesson for anyone repeating this work.

Exactly one debarred firm genuinely remains in the 2026 payroll data: **EGL Drainage Solutions & Repairs**, debarred July 17, 2025 for prevailing-wage violations under a Nassau County fiscal officer. The screen — with no knowledge of the debarment — independently flagged it: 33 findings in Nassau County, including steamfitter HVAC work paid \$32.40–\$40.00 against a scheduled \$46.60 floor and insulation work paid \$45.39–\$47.80 against \$71.76, every cited rate hand-verified

in the schedule PDF. All 33 findings survive the Tier-2 screen. On the screenable adjudicated population the screen’s recall is 1 of 1 — evidence that the screen points at the behavior regulators actually punish, not a statistical study ($n = 1$). EGL’s 2026 payroll lines postdate its debarment effective date; because debarment bars new awards while work on pre-existing contracts may lawfully continue, we do not characterize the lines as debarment violations.

4.2 False-positive controls and hand verification

The pilot slice contained two known traps, and the screen handles both: a line paid \$32.33 equal to the scheduled demolition-laborer rate is not flagged, and a line paid \$22.87 equal to a published first-term apprentice rate is caught by the apprentice screen. Beyond the controls, the floors behind every named example in this paper and on the companion site were verified by hand against the cited effective-dated schedule PDF — including the largest-exposure candidate (heavy & highway laborer and operating-engineer floors of \$42.93 and \$53.82 in Schoharie County) and three alias-derived floors.

5 The standing candidate set

The standing set has structure that random data-entry noise would not produce:

- **Flat round rates across counties.** The second-ranked contractor shows 237 standing high-confidence lines across five counties with workers filed as heavy/highway operating engineers and electricians at a flat \$20.00 per hour against floors near \$50 — a payroll practice, not a typo.
- **Large persistent gaps in licensed trades.** Electrical contractors paid \$16.00–\$16.50 against county floors above \$51 (68% short), sustained across months.
- **Concentration.** 468 of 3,503 contractors carry any standing high-confidence finding; 20 account for 45.8% of standing exposure. The screen’s purpose — turning 588,505 lines into a short, ranked review queue — is met.

Each candidate ships with: the public wage figure, the cited effective-dated schedule rate and its provenance (including alias tags where applied), the Tier-2 verdict, and the caveat set of Section 6. Named contractors are identified *only* as screening candidates from public records requiring investigation. Firms with adjudicated records (debarment) are identified as adjudicated fact, with dates.

6 Limitations

1. **Cash wage only.** The public dataset omits supplements. New York treats the wage and supplement obligations separately, so cash materially below the scheduled wage is a genuine investigation trigger — but no finding here speaks to total compensation.
2. **Apprentice registration is not publicly verifiable.** The apprentice screen removes rate-matched lines from priority; conversely, a line that matches no apprentice rate could still belong to a registered apprentice under an unpublished arrangement. Only the Department’s registration records resolve this.
3. **The floor is deliberately minimal.** Using the family-minimum rate under-flags: sub-classification-specific underpayment (paid the lowest group’s rate for higher-group work) is

invisible here. Precision was bought at the cost of recall.

4. **A certified payroll states the classification paid, not the work performed.** Misclassification — the largest enforcement category in practice — cannot be proven from this data and is out of scope.
5. **County is derived.** 99.8% coverage with per-line provenance; genuinely multi-county project filings are excluded rather than guessed.
6. **The exposure figure is indicative.** Shortfall $\times$ straight-time hours on standing findings; it is a prioritization weight, not a damages estimate.

7 Recommendations

For enforcement use. The standing set is a ranked review queue: each of the top candidates can be resolved by one records pull (the full certified payroll, which the contractor already filed) plus an apprentice-registration lookup — both routine for the Department, neither public.

For the data publisher. Four fields would transform what public screening can do, at near-zero marginal cost since all exist in the filed payrolls: (1) supplements paid, (2) a reliable apprentice indicator with term, (3) project county, and (4) the applicable schedule classification as filed. Absent those, this paper documents the workarounds — and their limits.

A Replicating the analysis for New York City

New York City is a distinct prevailing-wage jurisdiction: the NYC Comptroller sets §220 schedules for city public work and receives certified payrolls under Labor Law §220-j / Local Law. A pilot of exactly this method was completed against the Comptroller’s data in July 2026; this appendix records what a full NYC deployment requires, and where NYC’s data is better and worse than the State’s.

A.1 Data required, and where it lives

1. **Certified payrolls.** The Comptroller operates a public certified-payroll directory. A public JSON listing endpoint enumerates filed payrolls (no authentication); each payroll’s worker-week line items are retrievable through the site’s SSRS ReportViewer export in XML, CSV, and PDF formats. The XML export is the reliable parse target; the PDF export is the human-readable evidence snapshot each finding should link to. **Crucially, and unlike the State dataset, the NYC line items include the supplement rate paid** (the fund/admin field), so NYC supports a *benefit-shorting* screen in addition to the cash-wage screen — the pilot ran both.
2. **Rate schedules.** The Comptroller publishes §220 schedule PDFs (Construction Worker; separate Building Service and Apprentice schedules). As with the State, the current edition is not sufficient: the schedule year runs July–June, so screening any prior-year work requires the *archived* prior edition. The pilot’s controlling result: naively citing current-year rates against early-2026 work manufactured 74 false underpayment signals on a corpus that was in fact fully compliant. Rates step mid-year within an edition; the library must store one row per effective period, not per classification.

3. **Validation ground truth.** The Comptroller’s employer-violations dashboard lists debarments (which mirror the State list) and settlement outcomes; the settlement charts are rendered client-side, so case-level capture requires either browser-level capture of the underlying JSON or a records request. This is the best available path to a genuine precision measurement in either jurisdiction.

A.2 Pipeline adaptation

The State pipeline carries over with three substitutions. Harvest: the listing endpoint plus SSRS export replaces the Socrata bulk download (slower per record; snapshot every export). County attribution: unnecessary — NYC schedules are citywide. Classification mapping: NYC payrolls state fine-grained classifications that map nearly one-to-one to schedule entries, so the State’s family-minimum convention is replaced by direct hand-verified mapping (the pilot verified each of its corpus classifications against the schedule’s scope descriptions); the apprentice screen keys on the Comptroller’s separate apprentice schedule.

A.3 Pilot result (July 2026)

The pilot screened the complete then-available corpus — 13 certified payrolls, 3 contractors, 94 worker-week lines (all Parks Department work, March–June 2026) — against both the wage and supplement floors in force on each week-ending date: zero underpayment findings, zero supplement-shorting findings, zero unmatched lines, and 74 false positives prevented by effective-date discipline. A small, compliant corpus, reported as such; the machinery is proven end-to-end and scales with the corpus. The State-side result above indicates what to expect as NYC coverage grows.

A.4 Epistemic posture

Identical to the State screen: findings are allegations that prioritize review; supplements and apprentice status resolve only against non-public records (in NYC, supplements are largely resolved by the published data itself); named parties appear only as screening candidates unless adjudicated.

Concourse Technologies — Research · July 2026

Data sources: data.ny.gov certified payrolls; NYSDOL prevailing-wage schedules and debarment list; NYC Comptroller §220 schedules and certified-payroll directory. All pipeline stages are reproducible; unresolved lines are logged, never dropped.