Methodology for International AI Compensation Data

United States, United Kingdom, and Canada (Tier 1)

Justin Bartak
with Claude (Opus 4.7, Anthropic) as drafting partner

Orbyt Intelligence — orbytjobs.ai/orbyt-intelligence/methodology

Methodology for International AI Compensation Data

United States, United Kingdom, and Canada (Tier 1)

A peer-reviewable preprint describing the data sources, occupation crosswalks, currency normalization, reconciliation rules, and confidence interval calculations used by Orbyt Intelligence for cross-jurisdictional AI compensation estimates.

Abstract

The AI compensation data market segments into three product categories: enterprise compensation survey products (Mercer, Aon Radford, Willis Towers Watson) at $25,000 to $100,000 per seat per year; consumer freemium products (Levels.fyi, Glassdoor, Payscale) with opaque methodology and US-dominant coverage; and payroll-feed products (Pave, OpenComp) at $50,000+ per company with individual-record live data. None of these occupies the developer-API segment with a transparent methodology contract across multiple jurisdictions.

This paper documents the methodology behind Orbyt Intelligence's Tier 1 international coverage: the United States, the United Kingdom, and Canada. Tier 1 reconciles eight government-sourced data inputs (BLS OES, DOL H-1B LCA, US state pay-transparency portals; ONS ASHE, HMRC PAYE RTI, UK Skilled Worker Visa going rates; ESDC Open Government wages, Statistics Canada Web Data Service) into per-country weighted estimates updated monthly. Reconciliation weights are locked per country and disclosed on every estimate. Currency normalization uses ECB-backed daily rates from the Frankfurter API. Cross-country sanity bounds quarantine implausible values (any non-US value greater than five times or less than one-fifth the US baseline). Every estimate carries its methodology version, source breakdown, sample size, confidence interval, and disagreement flag in the locked response envelope.

The paper makes three contributions. First, the published cross-jurisdictional reconciliation methodology with explicit per-country weights, the locked 25% disagreement threshold, and the 5× cross-country sanity bound. Second, the open-source Orbyt AI Role Taxonomy (3,445 roles across 13 hubs, 687 AI-classified, CC BY 4.0). Third, the queryable per-data-point provenance trail accessible at `/api/v1/intelligence/lineage/[data_point_id]`, exposing every source's contribution and weight to every reconciled estimate.

The paper is honest about limitations. No official government concordance exists between US SOC 2018 and UK SOC 2020, or between NOC 2021 and UK SOC 2020. Orbyt's mappings for these pairs are hand-curated and tagged as `expert-judgement` confidence. Tier 1 coverage is metro-level, not neighborhood-level. Survey-window-to- publication lag means published values are backward-looking floors for AI roles, with the `/projections` endpoint providing forward curves. Community-submission Tier C data carries selection bias that the privacy floor mitigates but does not eliminate.

The Role Taxonomy is published under CC BY 4.0 for any third-party adoption. The methodology paper accepts pull requests at the Orbyt GitHub repository. The API is priced at $99 to $4,999 per month across four tiers, accessible via REST, CLI, and the Model Context Protocol. SDK auto-generates from the OpenAPI specification this paper references.

---

1. Introduction

AI labor data is a market with three established product categories, none of which serves the developer who needs a programmatic, transparent, multi-jurisdictional source of compensation truth.

This paper describes a fourth category and the methodology that makes it possible: the developer-API for AI compensation data, priced like Stripe, accessed via REST, the Model Context Protocol, and a command- line client, with Bloomberg-grade source attribution on every value.

The first draft of this paper is published as a preprint under CC BY 4.0. The repository at orbytjobs/orbyt accepts pull requests against the source markdown. The methodology version string in the API response bumps when the methodology changes; the lineage table preserves every prior version forever so historical observations remain queryable in their original methodology.

1.1 The state of AI labor data

Three product categories define the existing AI compensation data market.

Enterprise compensation survey products anchor the top of the market. Mercer, Aon Radford, and Willis Towers Watson sell annual benchmark surveys at $25,000 to $100,000 per seat per year. Sales motion is procurement-gated; access is consultant-mediated. Coverage spans global jurisdictions and granular role taxonomies. Methodology is documented in deliverable PDFs that the buyer receives after contract execution. Customers are HR analytics teams at large employers, executive compensation committees, and benefits brokers.

Consumer compensation data products anchor the middle of the market. Levels.fyi, Glassdoor, and Payscale offer freemium access to crowd-sourced salary disclosures. Coverage is US-dominant and the international layer is thin. Methodology is partially documented on each product's site but the underlying selection mechanism (who chooses to self-disclose) introduces bias that the products acknowledge inconsistently. Customers are job seekers comparing themselves to their peers; secondary customers are recruiters using the consumer surface as a sourcing input.

Payroll-feed compensation products anchor the third category. Pave and OpenComp sell individual-record live data from direct HR-systems integrations at $50,000+ per company per year. Coverage is US-only or US-dominant. Methodology rests on the assumption that the connected companies are representative, which they often are not (Pave's network skews toward VC-backed scaleups; OpenComp's toward enterprise-tier mid-market). Customers are HR teams at the member companies and a small number of external buyers paying for benchmark access.

These three categories collectively serve buyers who fit the implied sales model: HR teams with budget for a five-figure subscription; job seekers willing to trade self-disclosure for free aggregate access; finance teams at scaleups happy to plug their payroll into a vendor's data lake.

Nobody serves the developer.

1.2 The gap in the developer-API segment

The developer building a compensation feature into a product needs different things than any of the three established categories provide.

The developer needs **programmatic access**, not a PDF report. The developer needs **transparent methodology**, not a black box, because the developer's product will be cited downstream and the citation chain matters. The developer needs **predictable pricing**, not a procurement RFP that takes 90 days to close. The developer needs **multi-jurisdictional coverage**, not US-only, because the developer's product will have customers outside the United States. The developer needs **a methodology version that doesn't change silently**, because the developer's product will lock to a specific version for reproducibility.

Stripe-style developer-API pricing matches what the developer expects: monthly billing, transparent tiers, self-serve signup, no contract minimums, an SDK in the developer's preferred language, generated from a published OpenAPI specification. The compensation data market has no incumbent product at that price point with that methodology contract.

Why has nobody built it? Three reasons we can identify.

Distribution economics. The enterprise compensation survey incumbents (Mercer, Aon, WTW) make their revenue from $25,000-plus contracts. A $99-per-month developer tier requires 250 customers to match one enterprise seat, with 250 times the support surface. The economic gravity pulls those incumbents away from the developer segment.

Methodology requirements. A developer-API customer needs reproducible numbers they can cite in their own product. Neither freemium consumer products (which deliberately obscure methodology to protect the consumer-grade brand) nor payroll-feed black boxes (which can't disclose methodology without revealing the connected network's composition) can offer that.

International scope. Every existing product in the AI compensation space is US-dominant. Building genuine multi-country coverage requires solving the cross-jurisdictional occupation crosswalk problem. The crosswalk problem is unsolved (this paper §3 documents the absence of official concordances) and requires deliberate engineering investment.

The gap is real. Orbyt Intelligence fills it.

1.3 This paper's contribution

This paper documents the methodology Orbyt Intelligence uses to publish AI compensation data across the United States, the United Kingdom, and Canada through a single locked API contract. The contributions are three.

First, the published cross-jurisdictional reconciliation methodology. Per-country source weights are documented explicitly in §4.1 (US: BLS 0.40, H-1B LCA 0.30, pay-transparency 0.30; UK: ASHE 0.60, HMRC RTI 0.20, Visa 0.20; Canada: ESDC 0.60, StatCan 0.40). The 25-percent disagreement threshold is locked in §4.3. The 5-times cross-country sanity bound is locked in §4.4. Every locked choice is documented with rationale so a third party can audit it and so the choice itself is reproducible.

Second, the open-source Orbyt AI Role Taxonomy. 3,445 roles across 13 hubs, 687 AI-classified, published as a standalone npm package at `packages/role-taxonomy/` under CC BY 4.0. The taxonomy is the structural data; the salary numbers are the paid product. Letting the taxonomy travel freely means every AI tool that adopts the slugs becomes a pointer back to Orbyt's data. §3.1 documents the package, the versioning rules, and the trade-off this open license represents.

Third, the queryable per-data-point provenance trail. Every estimate Orbyt publishes carries a `data_point_id` that resolves to a row in the `data_lineage` table accessible via `/api/v1/intelligence/lineage/[data_point_id]`. The lineage row lists every source that contributed, the weight each source carried, the value each source provided before weighting, the reconciled output, and the methodology version that produced it. A customer who quotes an Orbyt number in their own product can link the citation to the data point, and a reader can verify the underlying calculation against the source agencies' published values. This is the Bloomberg-grade transparency the developer-API customer needs.

The rest of this paper documents the engineering and statistical choices that make those three contributions possible. The Tier 1 international launch is the first published methodology version that covers more than one country. Phase B (additional countries) will ship under a separate RFC with appropriate version-string bumps.

---

2. Data sources

Orbyt Intelligence's Tier 1 international coverage draws from eight distinct sources across three countries. Five are direct publications from national statistical agencies (BLS, ONS, HMRC, ESDC, StatCan). Two are government regulatory filings (DOL H-1B LCA, UK Skilled Worker Visa). One is state-level pay-transparency portals (CA / CO / NY / WA). Every source is free, government-issued, and refreshes on a published schedule. None require a paid API key. The methodology paper publishes the source URL for every ingested observation so a third party can independently verify the value.

2.1 United States

The US baseline is established at Phase 2A and predates this paper. Three sources reconcile at locked weights: BLS OES at 0.40, DOL H-1B LCA at 0.30, and state pay-transparency portals at 0.30. Per-country weights sum to 1.0; the methodology paper documents the rationale.

2.1.1 BLS Occupational Employment Statistics (OES)

Annual publication from the US Bureau of Labor Statistics. The May release each year covers approximately 830 occupations across 388 metropolitan statistical areas (CBSAs). Each cell publishes a wage distribution: 10th, 25th, 50th (median), 75th, and 90th percentiles.

BLS suppresses cells where the sample size falls below an internal disclosure threshold. Suppressed values are marked with `*` (no estimate available) or `**` (employment-level suppressed). The connector at `app/api/v1/intelligence/_ingestion/sources/bls-oes.ts` skips suppressed rows silently per the BLS disclosure policy.

Source URL pattern (Phase 0 verified): `https://www.bls.gov/oes/special-requests/oesm{YY}all.zip` where `{YY}` is the two-digit release year. The download is a ZIP containing an Excel workbook. Orbyt's operator-side workflow extracts the workbook, converts to CSV, and hosts the CSV at a stable URL via the `BLS_OES_DATA_URL` environment variable. The connector parses CSV input directly; binary XLSX parsing is intentionally out of scope to keep dependencies light. Source weight in US reconciliation: 0.40.

2.1.2 DOL H-1B Labor Condition Application disclosures

Quarterly publication from the US Department of Labor, Office of Foreign Labor Certification. Wage levels declared by US employers filing Labor Condition Applications for H-1B visa sponsorship. The dataset has structural skew toward technical occupations (the H-1B program disproportionately covers software engineers, data scientists, hardware engineers, and research roles), which means the technical- occupation cells in H-1B LCA carry higher effective sample sizes than the BLS OES equivalent cells for the same occupation. This is desirable for AI compensation data specifically; H-1B's overrepresentation of the roles we care about is a feature, not a bug.

The connector at `app/api/v1/intelligence/_ingestion/sources/h1b-lca.ts` reads operator-extracted CSV. Source weight: 0.30.

2.1.3 State pay-transparency portals

Pay-transparency laws in California (effective 2023), Colorado (2021), New York (2023), and Washington (2023) mandate that employers post salary ranges on every job listing. Job listings are scraped at the state portal level (where available) or via the COMMON CRAWL CC0 archive (which captures publicly-listed pages). The connector at `app/api/v1/intelligence/_ingestion/sources/pay-transparency.ts` and its CommonCrawl variant at `app/api/v1/intelligence/_ingestion/sources/commoncrawl.ts` aggregate salary ranges from these listings. Sample sizes vary by state and occupation; states that mandate disclosure aggressively (California, New York) produce denser samples than states with narrower mandates (Washington's $100K+ threshold). Source weight: 0.30 combined across all states.

2.2 United Kingdom

The UK reconciliation bucket sums to 1.0 across three sources: ONS ASHE at 0.60, HMRC PAYE RTI at 0.20, and the Home Office Skilled Worker Visa going-rate table at 0.20. ASHE is the primary occupation- coded source; HMRC RTI provides a monthly geographic-level signal that contributes to every role in a given geography; the Visa going-rate table constrains the lower bound only.

2.2.1 ONS Annual Survey of Hours and Earnings (ASHE)

Annual publication from the Office for National Statistics. The October release each year reports gross weekly earnings by SOC 2020 occupation code at two granularity levels: Table 2 publishes at two-digit SOC granularity for the full UK, country, and region breakdown; Tables 14 through 16 publish at four-digit SOC granularity with higher suppression rates in smaller regions.

Each cell publishes median, 25th and 75th percentile gross weekly pay, plus the number of jobs (in thousands) backing the estimate. Orbyt's connector at `app/api/v1/intelligence/_ingestion/sources/uk-ons-ashe.ts` parses both Table 2 and Tables 14-16 from operator-extracted CSV input. Weekly pay multiplies by 52 weeks to produce an annual figure; the ASHE methodology paper notes this is a slight overstatement compared to a strict calendar-year approach but is the convention used by ONS in their own published derived statistics.

Source URL pattern (Phase 0 verified): the ONS Open Data site at `https://www.ons.gov.uk/file?uri=/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/occupation2digitsocashetable2/{year}provisional/ashe` for Table 2; analogous paths for Tables 14-16. The ZIP contains an XLS file. Same operator-side extraction pattern as BLS OES.

Source weight in UK reconciliation: 0.60. This is the heaviest weight because ASHE is the most directly comparable to BLS OES in shape and methodology.

2.2.2 HMRC PAYE Real Time Information

Joint publication from HM Revenue and Customs and the Office for National Statistics. Monthly bulletin. Reports median gross monthly pay by UK / NUTS 1-3 (regional) / local authority levels. Phase 0 verification (2026-05-17) confirmed an important property of this source: PAYE RTI is geographic-only. It carries no SOC occupation breakdown anywhere in the published file. The HMRC + ONS classification of every employed UK individual into the PAYE system produces an aggregate at the geographic level, not at the occupational level.

Orbyt's reconciliation handles this by fanning out the geographic signal: one PAYE RTI row (e.g., London median monthly pay of 3,500 GBP) becomes N entries with role_id = each Orbyt role in the applicableRoles list, all sharing the same value, at base_median field only. Per RFC-006 §8.4, this contributes a soft drag at 0.20 weight toward the geography's population-mean monthly pay. The methodology paper notes the duplication explicitly: a reader inspecting the lineage trail for a London role will see HMRC RTI listed once as a contributor, with sample_size reflecting the full PAYE employment count. That sample size is one observation, not N independent observations.

Source URL pattern: `https://www.ons.gov.uk/file?uri=/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/realtimeinformationstatisticsreferencetableseasonally{year}.xlsx` .

Source weight: 0.20.

2.2.3 UK Skilled Worker Visa eligible occupations

The UK Home Office publishes Appendix Skilled Occupations of the Immigration Rules, which sets the minimum salary an employer must pay a Skilled Worker visa holder for each eligible SOC 2020 occupation code. The page is HTML-only on gov.uk; there is no CSV or spreadsheet download. The connector at `app/api/v1/intelligence/_ingestion/sources/uk-visa-occupations.ts` uses cheerio to parse the HTML table, extracting four-digit SOC 2020 codes and their associated going-rate salaries in GBP per year.

By design, going rates are legal floors. The UK government sets them below market clearing prices to allow employers to legitimately hire visa holders at the bottom of the market band. Orbyt's connector emits only `base_low` from this source, constraining the lower bound of UK reconciliation without dragging the median.

Source URL: `https://www.gov.uk/government/publications/skilled-worker-visa-eligible-occupations/skilled-worker-visa-eligible-occupations-and-codes` .

Source weight: 0.20.

2.3 Canada

The Canada reconciliation bucket sums to 1.0 across two sources: ESDC Open Government wage data at 0.60 and Statistics Canada Web Data Service at 0.40. Geographic coverage differs between the two: Open Government covers Economic Region granularity, StatCan WDS covers national headline (province / CMA breakdown lands in a follow-up). The methodology paper documents the consequence: today's reconciled Canadian estimates blend the two sources only when both fire for the same (role, city) tuple; otherwise each source reconciles alone with 1.0 weight in its own geographic slice.

2.3.1 ESDC Open Government Wage Data

Annual November release from Employment and Social Development Canada via Canada's Open Government Portal. Hourly low, median, and high wages by NOC 2021 five-digit occupation code at Economic Region (four-digit) and provincial granularity. The CSV downloads directly from a stable URL on the Open Government Portal at approximately 200KB total.

The connector at `app/api/v1/intelligence/_ingestion/sources/ca-opengov-wages.ts` parses the CSV using a synonym-tolerant header map (ESDC has renamed columns across releases; the synonym map captures the variants). Hourly wages multiply by 2,080 hours (40 hours × 52 weeks) to produce an annual figure.

Source URL: `https://open.canada.ca/data/dataset/adad580f-76b0-4502-bd05-20c125de9116/resource/9da94d63-b178-4a64-aeb3-b6a3bd721ad2/download/2a71-das-wage2025opendata-esdc-all-19nov2025-vf.csv` .

Source weight in CA reconciliation: 0.60.

2.3.2 Statistics Canada Web Data Service

Monthly publication from Statistics Canada via the WDS REST API. The Labour Force Survey (LFS) Table 14-10-0064-01 publishes median hourly wage rate by NOC 2021 occupation. The Survey of Employment, Payrolls and Hours (SEPH) publishes complementary employment counts.

WDS keys time series by V-prefixed numeric vector codes. To query the median hourly wage rate for NOC 21231 (software engineers and designers), the connector queries `vectorIds=[<numeric>]` . The mapping from NOC code to vector code is not directly published as a CSV; it must be recovered from the `getCubeMetadata` endpoint that returns the cube's dimensional structure. RFC-006 §8.4 specifies this as a one-time-per-year bootstrap that the operator runs after each annual cube revision. The bootstrap cron at `/api/cron/statcan-vector-bootstrap` returns the cube metadata for operator review; the operator updates the seed map in code and ships a PR.

The connector at `app/api/v1/intelligence/_ingestion/sources/ca-statcan-wds.ts` emits one SourceEntry per (Orbyt role, base_median) tuple per vector. Hourly wage rate multiplies by 2,080 hours for annual.

Source URL (POST endpoint): `https://www150.statcan.gc.ca/t1/wds/rest/getDataFromVectorsAndLatestNPeriods` .

Source weight: 0.40.

2.4 Source reliability classification

The methodology paper sorts sources into three reliability tiers per RFC-006 §5.3:

Tier I — Direct government statistical agency. BLS OES, ONS ASHE, StatCan WDS, ESDC Open Government wages. Each is a national statistical agency's own published estimate. The agency's own methodology paper is the definitive reference for the source's sampling design, weights, suppression rules, and confidence intervals. Revisions happen on annual cycles. Orbyt's connector preserves the agency's published values verbatim.

Tier II — Government regulatory filings. DOL H-1B LCA, UK Skilled Worker Visa going rates. Government-issued but filed by employers in a regulatory process. Suitable for bounding the reconciliation rather than centering it. H-1B LCA's wage levels are employer-declared and skew toward the upper band of legitimate compensation for the occupation (employers minimize visa-rejection risk by overstating). UK Visa going rates are intentional legal floors set below market clearing prices. Orbyt's connectors emit appropriate fields only (LCA contributes to base_low/median/high; Visa contributes to base_low only).

Tier III — State pay-transparency portals. California, Colorado, New York, Washington. Mandatory disclosure of salary ranges on posted job listings. Methodology varies by state. Sample selection is biased toward roles employers actually post (high-demand roles overrepresented; senior roles underrepresented because senior hiring frequently happens through executive search where the disclosure mandate is weakly enforced).

2.5 Update cadence

Per-source refresh schedules and the implications for data freshness:

Source	Cadence	Lag at first publication
BLS OES	Annual (May)	~6 months behind survey window
DOL H-1B LCA	Quarterly	~3 months
Pay-transparency	Continuous (scraped weekly)	< 1 week
UK ONS ASHE	Annual (October)	~6 months
UK HMRC PAYE RTI	Monthly (19th of month)	~1 month
UK Skilled Worker Visa	When Home Office updates Appendix	Variable
CA ESDC Open Gov	Annual (November)	~6 months
CA StatCan LFS	Monthly	~6 weeks
Frankfurter FX	Daily (ECB business days)	< 1 day

Each Orbyt monthly ingestion run at 04:00 UTC refreshes whatever upstream sources have new data. When a source is unavailable for a given cron tick (HTTP 404, schema drift, FX outage), the connector records a failure in `IngestionRunResult.sources_failed`; reconciliation continues with the remaining sources, and the country's effective sample size declines for that tick. Last-known reconciled values stay in the canonical dataset until the next successful tick produces fresh values. No customer-facing outage from any single source failure.

---

3. Role taxonomy and occupation crosswalk

The role-to-occupation mapping is the structural backbone of cross- jurisdictional reconciliation. Without a consistent way to translate "AI Engineer" or "Data Scientist" across US SOC 2018, UK SOC 2020, and Canadian NOC 2021, no two countries' wage data can be combined. The mapping work is not glamorous. It is the load-bearing element of the methodology. This section documents what Orbyt built, what information loss is unavoidable, and how the lineage trail preserves honesty about which mappings are direct and which are expert judgment.

3.1 The Orbyt AI Role Taxonomy

Orbyt maintains an open-source role taxonomy at `packages/role-taxonomy/`, licensed CC BY 4.0. As of the 2026.4 methodology release, the taxonomy covers 3,445 distinct roles organized into 13 thematic hubs (Engineering, AI/ML, Data, Design, Product, Operations, Sales/Marketing, Security, Finance, Legal, People, Customer Success, Healthcare). 687 of those roles are classified as AI-related by the hub assignment heuristic; the AI subset is the dense core of this paper's coverage.

The taxonomy is generated by `scripts/export-role-taxonomy.ts` from two curated source files: `public/data/salary-roles-prose.json` and `app/(marketing)/salaries/salary-hubs.ts`. The export script produces three JSON artifacts in the published package: a roles file, a hubs file, and a manifest with generation timestamps and per-hub role counts. The package's published versioning rule is locked: role slugs are sacred. A slug published in one release never changes meaning in a future release. New roles ship as MINOR version bumps. Hub re-classifications ship as MAJOR version bumps with full CHANGELOG context.

The taxonomy carries only structural metadata: slug, display title, hub assignment, BLS SOC code (where mapped). It deliberately omits the commercially-valuable data Orbyt's paid product provides: salary numbers, salary drivers, career ladders, equity bonus signing-bonus medians, remote and company-size multipliers. The structural data is open; the data product is paid. This split lets the taxonomy travel into any third-party tool that needs canonical role identifiers without giving away the commercial layer.

3.2 The crosswalk problem

Tier 1 international coverage requires reconciling data published in three different occupation classification systems:

- **United States: SOC 2018.** Standard Occupational Classification, 2018 revision. Used by BLS OES, DOL H-1B LCA, and most US federal labor statistics.
- **United Kingdom: SOC 2020.** Standard Occupational Classification
2020. Used by ONS ASHE, HMRC PAYE RTI, and UK Home Office Skilled Worker Visa eligibility lists.
- **Canada: NOC 2021.** National Occupational Classification 2021, the latest revision published by Statistics Canada and Employment and Social Development Canada.

Phase 0 verification (2026-05-17) examined the published concordances between these three systems. The finding is unambiguous: no direct official concordance exists between any pair.

Statistics Canada publishes a NOC 2016 → US SOC 2018 concordance, which is partially usable but deprecated as Canada has since moved to NOC 2021. NOC 2021 → US SOC 2018 requires chaining through NOC 2021 → NOC 2016 → US SOC 2018, with information loss at each step because the chained classifications are themselves many-to-many.

US SOC 2018 ↔ UK SOC 2020 has no official concordance. Period. The US Bureau of Labor Statistics and the UK Office for National Statistics maintain their classifications independently with no joint mapping work.

NOC 2021 ↔ UK SOC 2020 has no official concordance either.

This is the honest baseline. Orbyt cannot promise the elegance of official-concordance-driven mapping. The mapping work is hand-curated, documented per-pair, and surfaced to API consumers via a confidence tag on every estimate.

3.3 Crosswalk confidence tiers

Per RFC-006 Appendix B, every (Orbyt role × country) mapping carries one of three confidence values, persisted alongside the mapping in the `role_occupation_codes` table:

`direct` . An official government concordance maps the source country's classification to the target country's classification. This is rare in the Tier 1 scope. The only application within Tier 1 is the US-internal SOC 2010 → SOC 2018 transition (which Orbyt absorbed in Phase 2A) and the partial NOC 2016 → US SOC 2018 concordance (which Tier 1 leverages where it produces sensible chains to NOC 2021).

`inferred-via-NOC2016` . Specific to Canada-to-US mappings. NOC 2021 codes are chained through NOC 2016 to US SOC 2018 by following StatCan's published NOC 2016 → SOC 2018 concordance, then back-mapping NOC 2021 → NOC 2016 via StatCan's published version- update table. Information loss accumulates at both steps because the underlying classifications are many-to-many. A single NOC 2021 code can map to multiple NOC 2016 codes, and each NOC 2016 code can map to multiple SOC 2018 codes. The methodology paper publishes the chained result with this confidence tag so the customer knows the mapping has compounded uncertainty.

`expert-judgement` . No concordance exists. Orbyt's owner curated the mapping using occupation titles, O*NET skill profiles, and labor- market familiarity. This applies to all US SOC 2018 ↔ UK SOC 2020 mappings and to all NOC 2021 ↔ UK SOC 2020 mappings. Expert-judgement mappings are the lowest-confidence tier; the methodology paper discloses this explicitly per role to set the customer's expectation. A reader who wants to use Orbyt's UK data needs to know that the connection between "Orbyt's AI Engineer slug" and "ONS ASHE's SOC 2425 Actuaries Economists Statisticians" is a judgment call, not a government-blessed equivalence.

3.4 The 60-mapping seed

The Tier 1 launch ships with 60 hand-curated mappings: 20 AI-relevant Orbyt role slugs × 3 countries = 60 (slug, country, occupation code) triples. Each triple carries:

- `occupation_system` : one of SOC_2018, SOC_2020, NOC_2021
- `occupation_code` : the country-specific code (e.g., `15-1252` , `2425` , `21211`)
- `occupation_title` : the source agency's title for that code
- `coverage_tier` : A, B, or C per §3.5
- `confidence` : `direct` , `inferred-via-NOC2016` , or `expert-judgement`
- `notes` : human-readable provenance reasoning

The 60 mappings are persisted to the `role_occupation_codes` table. When the connector layer queries the canonical mapping at fetch time, it joins on (role_slug, country_code) and pulls the codes for that country. The mapping is the source of truth for the connector's NOC → role and SOC → role lookups.

The seed publishes alongside this paper as a CC BY 4.0 dataset. Phase B expansion will grow the mapping table; the methodology paper will publish each expansion's deltas in a versioned CHANGELOG.

3.5 Coverage tier per role per country

Beyond the confidence tag on the mapping itself, each role × country pair gets a coverage tier per RFC-006 §6.1 reflecting the quality of the underlying data, not just the mapping:

Tier A — direct government data. A specific occupation code maps 1:1 (or with high coverage) to the role and the underlying source publishes data for that code. Orbyt's `software-engineer` slug maps to BLS SOC `15-1252` (US, Tier A), to ONS SOC `2136` (UK, Tier A), and to NOC `21231` (CA, Tier A). All three sources publish data for the mapped code with sufficient sample size to clear the disclosure floor in major metros.

Tier B — inferred from related code. The role's mapping points to a parent or sibling occupation code; Orbyt reports the parent's data with a tier-B confidence flag. For example, a role like `ml-research-scientist` doesn't have a dedicated SOC 2018 code distinct from `data-scientist` ; both map to BLS SOC `15-1252` . Reporting `ml-research-scientist` 's estimate as the BLS SOC `15-1252` estimate is honest but coarse; the customer sees Tier B and understands the value reflects the broader code's median.

Tier C — community / baseline. The role's underlying source data falls below the privacy floor, or no usable government source exists. Tier C falls back to crowd-sourced submissions (Orbyt's `/api/v1/intelligence/salaries/crowd` endpoint, when sufficient samples accrue) plus the US baseline scaled by purchasing-power- adjusted FX. Tier C estimates are explicitly flagged in the API response so customers can decide whether to trust them.

The coverage_tier field is exposed on every Tier 1 API response per RFC-006 §9.2. A customer querying `/calculate?role=ai-engineer&city=london-uk&country=GB` receives `coverage_tier: "A"` or `"B"` or `"C"` alongside the salary estimate. This is the load-bearing transparency mechanism: the customer can trust the API to tell the truth about how confident they should be in any given number.

---

4. Reconciliation methodology

Multi-source data integration faces a fundamental tension: each source publishes its own estimate with its own sampling design, its own methodology, its own biases. A naive average across sources mixes strong signals with weak ones at equal weight. A purist single-source approach loses the cross-validation benefit of triangulation.

Orbyt's reconciliation engine resolves this with locked per-country weighted averaging, partial-source renormalization, 25%-deviation disagreement detection, cross-country sanity bounds, and methodology version stamping. The full implementation is at `app/api/v1/intelligence/_ingestion/reconcile.ts` . Every reconciled value is reproducible from the per-source contributions in the lineage record.

4.1 The weighted average

Each country has a locked weight set per RFC-006 §8.3:

Country	Source	Weight
US	BLS OES	0.40
US	H-1B LCA	0.30
US	Pay-transparency	0.30
UK	ONS ASHE	0.60
UK	HMRC PAYE RTI	0.20
UK	Skilled Worker Visa going-rate	0.20
CA	ESDC Open Gov	0.60
CA	StatCan WDS	0.40

Per-country weights sum to 1.0 independently. The reconciliation engine never blends across countries: a UK measurement and a CA measurement for the same (role, city, field) tuple land in separate buckets and reconcile with their own weight sets. The orchestrator at `app/api/v1/intelligence/_ingestion/orchestrator.ts` groups SourceEntries by the five-part tuple (role, city, segment, field, country) before invoking the engine.

The locked weights reflect Orbyt's reading of each source's methodological quality. BLS OES gets 0.40 because it is the most comprehensive US occupation survey, with the broadest geographic coverage and the most stable methodology. H-1B LCA gets 0.30 because its sampling bias toward technical occupations boosts the sample size of cells we care about for AI compensation, but the upward skew from employer self-declaration warrants discount. Pay-transparency portals get 0.30 with comparable rationale around the disclosure-mandate selection bias. The UK and Canada weights similarly balance the primary occupation-coded source against the secondary signals.

The weights are not derived from a formal Bayesian posterior. They are a locked deliberative choice, documented in this paper for transparency. Any future revision requires a new RFC and a methodology version bump.

4.2 Renormalization on partial source sets

When fewer than all of a country's sources fire for a given tuple, weights renormalize over the present set. This preserves the property that source_breakdown always sums to 1.0 (RFC-001 invariant).

For example, consider a US tuple where BLS OES and H-1B LCA both contribute but pay-transparency has no data for the region. The original weights are 0.40 / 0.30 / 0.30. The renormalized weights across the two present sources are:

- BLS OES: 0.40 / (0.40 + 0.30) ≈ 0.571
- H-1B LCA: 0.30 / 0.70 ≈ 0.429

Same formula applies to UK and CA partial sets. The reconciliation engine handles this via the `presentSources` cache in `reconcile.ts:reconcile()`.

A single-source case is the degenerate limit: source_breakdown is `{<source>: 1.0}`. Disagreement detection trivially returns false because there is no other source to disagree with.

4.3 Disagreement detection

OWNER-DECISIONS-2026-05-10 Q2 locked a uniform 25% threshold across all fields. When a multi-source reconciliation includes any single source whose value differs from the reconciled consensus by more than 25%, the reconciled value carries `disagreement_flag: true`.

The customer-facing API exposes this flag on every response that includes an Estimate. The /lineage endpoint at `/api/v1/intelligence/lineage/[data_point_id]` surfaces the per-source contributions so an analyst can identify which source diverged and investigate.

A flagged disagreement does not block the value from entering the canonical dataset. It is a transparency signal, not a quarantine trigger. Reconciliation produces a usable answer even when sources disagree; the customer is informed and can choose how to use the data.

Per RFC §9.1's locked Estimate shape, the disagreement_flag is a permanent v1 field. Future methodology revisions can tune the 25% threshold (and would bump the methodology version), but the field itself never disappears.

4.4 Cross-country sanity bounds

Reconciliation can produce an implausible value when a connector misparses a row (decimal-as-thousands, units mismatch, hourly vs annual confusion) or when an FX rate is stale by orders of magnitude. RFC-006 §8.3.2 specifies cross-country sanity bounds to catch these at ingestion time.

The check is post-reconciliation. After every reconciled tuple is computed for the run, the orchestrator builds a US-baseline cache keyed by (role, field). For each non-US reconciled value, it computes the ratio against the US baseline:

```
ratio = non_us_value / us_baseline_value
```

If `ratio > 5` or `ratio < 1/5`, the row is quarantined with the locked `value_outside_range` flag and does not enter the canonical dataset. The 5× factor is wide by design: London salaries 50% lower or 2× higher than San Francisco are entirely plausible (different markets, different cost-of-living, different industry mix). 5× deviation in either direction is implausible enough to indicate a data error, an FX bug, or a connector schema drift.

When no US baseline exists (Orbyt covers the role in UK or CA data but not in US data), the sanity check passes through unchecked. The quarantine bucket does not exist to constrain coverage where the US data simply does not match; it exists to flag values that conflict with the US data we do have.

The full implementation lives in `app/api/v1/intelligence/_ingestion/sanity-bounds.ts`. Boundaries are inclusive (a ratio of exactly 5 passes; exactly 1/5 passes).

4.5 Methodology version stamping

Every reconciled value carries a `methodology_version` string per RFC-001 §2.4. US reconciled values use `2026.2` (the Phase 2A baseline established before RFC-006). UK and Canada reconciled values use `2026.4-tier1-gold` (the locked Tier 1 tag).

The methodology version is exposed on every API response in the `attribution.methodology_version` field and on every lineage row. Customers who pin to a specific methodology version (rare) get a clear signal when the version bumps. Customers using the latest version automatically get the new methodology on the next refresh cycle.

When this paper's methodology changes in any material way (new source added, weights revised, disagreement threshold tuned, sanity bounds adjusted), the version string bumps in lockstep. The version history is documented in this paper's appendix; the lineage table preserves all prior versions so historical observations remain queryable in their original methodology version forever.

---

5. Currency normalization

Cross-jurisdictional reconciliation requires a common currency. The US baseline is USD-denominated by the US sources themselves. UK sources publish in GBP. Canadian sources publish in CAD (StatCan WDS) or hourly CAD wages (ESDC Open Government). Orbyt converts everything to USD inside the reconciliation engine while preserving the original native-currency value alongside the converted USD value on every observation, so downstream API consumers can surface either at query time.

5.1 The fx_rates table

Daily exchange rates are stored in the `fx_rates` table per migration 084. Each row carries the currency_code (ISO 4217), rate_date (YYYY-MM-DD), usd_rate (numeric), and source. The table is populated by the FX cron at `/api/cron/fx-refresh` running at 02:00 UTC daily, two hours before the monthly ingestion cron at 04:00 UTC. The lead time ensures every ingestion run has access to the morning's fresh rates.

The FX source is the Frankfurter API at api.frankfurter.dev. Frankfurter is an open-source aggregator backed by European Central Bank reference rates. Phase 0 verification (2026-05-17) replaced the original RFC plan that targeted exchangerate.host, which migrated to a paid-key access model. Frankfurter remains free, no key required, no rate limit at Orbyt's scale. The API returns rates in the USD → native direction (1 USD = 1.3756 CAD). The connector inverts to produce the native → USD direction Orbyt's salary math needs (1 CAD = 0.727 USD).

5.2 Native and USD on every observation

Per RFC-006 §9.2, every non-US SourceEntry carries five FX-related fields:

- `value` : USD-normalized, the canonical reconciliation input
- `value_native` : the native-currency value before FX conversion
- `currency_code` : ISO 4217 native currency
- `fx_rate_used` : the rate applied (native → USD)
- `fx_rate_date` : the rate's as-of date

The reconciliation engine operates on USD values for cross-source comparability. The orchestrator persists both currencies to the lineage row's `computed_value` block so the API can surface either at query time. A customer querying with `?country=GB` receives a response that includes both the GBP value and the USD value with the FX rate disclosed explicitly. The customer chooses which currency to display.

US SourceEntries omit all five fields. Per the locked types, the absence of a country_code defaults to 'US'; the absence of FX fields implies the value is already in USD. This back-compat preserves the pre-RFC-006 shape for every US-only API consumer.

5.3 Weekend FX handling

The European Central Bank does not publish weekend rates. Frankfurter returns Friday's rate when queried on Saturday or Sunday. The FX cron tolerates this: when the response's `date` field matches Friday and the query was made on a weekend, the upsert stamps the row with Friday's date. Subsequent ingestion runs querying for the latest rate retrieve Friday's value with Friday's date until the next business day's rate publishes.

This is correct behavior. Weekend FX volatility is structurally low (banks are closed; the rate fundamentally cannot move). The `as_of` field on the resulting salary observation accurately reflects which business day's exchange rate was applied.

5.4 FX outage policy

When Frankfurter is unavailable for a given daily cron, the FX cron records a failure and continues. The next 04:00 UTC ingestion cron uses the most recent successful rate from the `fx_rates` table. Daily FX volatility is typically less than 1% day-over-day; using a one-day-old rate produces a less-than-1% error in the USD-converted salary value, which is below the noise floor of the underlying salary signal (which itself updates monthly or annually).

The connector layer treats FX absence as a graceful-degrade case via the `resolveFxRate` helper. When the FX provider returns null and no static fallback is configured, the connector emits zero entries for that run, the orchestrator records the source as `succeeded-with-zero-rows` , and the country's effective sample size declines for that tick. Last-known reconciled values remain in the canonical dataset until the next successful cycle.

For extended outages (>24 hours), the operations runbook at `docs/operations/tier1-international-runbook.md` Procedure 3 documents the manual upsert path: an operator can record a row in `fx_rates` from any ECB-backed source (Bank of Canada, Bank of England, or any other ECB cross-rate publisher) to bridge the gap.

---

6. Sample size and confidence

Sample size and confidence interval reporting is the place where most compensation data products lose credibility. The freemium consumer products (Levels, Glassdoor) often report a median with no sample size at all. The enterprise comp survey products (Mercer, Aon Radford) report sample size in the back of a $25,000-per-seat report that few buyers actually read. The payroll-feed products (Pave, OpenComp) have real sample sizes but black-box methodology around how those samples combine.

Orbyt's commitment is to publish the sample size on every estimate and to be explicit about the cases where the published number doesn't reflect what a statistician would call "independent observations." This section is the honest disclosure work. Some of the structural choices below look ugly when stated openly; they are nonetheless correct, and customers benefit from knowing them.

6.1 The disclosure floor

Tuples with combined sample size below 5 are quarantined per `SAMPLE_SIZE_FLOOR` in `app/api/v1/intelligence/_ingestion/orchestrator.ts`. Quarantined tuples do not enter the canonical dataset. The customer-facing API never sees them. The lineage table never sees them.

The threshold of 5 matches the threshold used by `/api/v1/intelligence/salaries/crowd` for individual-level submissions. It exists to protect against re-identification: a salary "median" computed from 2 or 3 observations is functionally close to publishing the individual values. Aggregating across at least 5 observations provides a meaningful privacy buffer.

For aggregated government sources (BLS OES, ONS ASHE, ESDC Open Government wages, StatCan WDS), the underlying agencies already suppress small-cell data per their own disclosure policies. Orbyt's floor of 5 is a secondary check; in practice, government-sourced cells that arrive at the connector have already passed the agency's own threshold (typically much higher than 5).

6.2 Per-country sample size practicality

The 5-sample floor rarely binds for US data because BLS OES survey sample sizes per CBSA × SOC cell run into the hundreds or thousands of observations for any common occupation. UK ASHE at Table 2 (two- digit SOC) sample sizes are typically in the tens of thousands per region; the floor is structurally unreachable. Canadian Open Government wages publish at the Economic Region level with sample sizes that reflect the underlying LFS survey design (~56,000 households nationally each month, distributed by population).

The floor binds in the practical edge cases:

- **Four-digit SOC at smaller UK regions.** ASHE Tables 14-16 with finer occupation granularity have higher suppression rates outside London + South East. Cardiff, Belfast, Glasgow occasionally produce cells under the floor for narrower occupation codes.
- **Smaller Canadian cities.** Halifax, Winnipeg, Quebec City have lower per-cell sample sizes than Toronto / Vancouver / Montreal. When all sources for a (role × city) tuple sum below 5, the tuple quarantines silently.
- **Community-submission Tier C data.** Until at least 5 submissions accumulate for a (role × city), Tier C falls back to the US baseline scaled by purchasing-power-adjusted FX. The methodology paper flags Tier C estimates explicitly on every response.

6.3 Sample size in HMRC RTI fan-out

This is the honest disclosure that this paper exists to publish.

UK HMRC PAYE Real Time Information is geographic-only. There is no SOC occupation breakdown anywhere in the published bulletin. One observation arrives at the connector: "median monthly gross pay in London for the reference month is 3,500 GBP." That single observation gets fanned out across every applicable role in London at normalize time, producing N entries with identical value and identical sample size (the full PAYE employment count, ~4.5 million for London).

The lineage row for any London role estimate will list HMRC RTI as a contributor, with sample_size: 4,500,000 (or similar). A reader unfamiliar with the fan-out pattern will reasonably interpret that as "HMRC sampled 4.5 million Londoners and produced an independent estimate for this specific role." This is not what happened.

What happened is: HMRC sampled 4.5 million Londoners and produced one number for all of London. Orbyt's connector replicated that single number across every Orbyt role slug operating in London. Each per-role estimate shares the same 4.5 million sample size, but those sample sizes are the same observation, not independent ones.

The 0.20 weight reflects this. HMRC RTI gets a small fraction of the UK reconciliation precisely because it is not occupation-specific. ASHE at 0.60 carries the per-occupation signal; HMRC at 0.20 contributes a geographic anchor toward the population mean. The arithmetic of the weighted average correctly handles the fan-out: HMRC pulls each London role's reconciled value toward London's population-mean monthly pay, capped at 0.20 weight.

A reader who wants to understand the true effective sample size for a specific London role's reconciled estimate should consult the lineage row's contributors list:

- ASHE entry's sample_size: the actual count of London respondents for the role's SOC code (typically thousands)
- HMRC entry's sample_size: 4.5 million, but representing the one-observation fan-out, not independent observations
- Visa entry's sample_size: 1, the legal-floor anchor

The mathematical convention for "effective sample size" in this multi-source weighted scheme is non-trivial; the methodology paper does not attempt a single-number summary because any such number would obscure the qualitative difference between independent observations (ASHE) and replicated geographic anchors (HMRC).

6.4 Sample size in StatCan WDS

StatCan WDS publishes wage rates per vector code, not per-vector sample counts. The connector emits `sample_size: 100` as a structural placeholder. The placeholder is intentionally chosen above the disclosure floor (5) to allow the data to reach reconciliation, but clearly below the order-of-magnitude that a real sample count would imply for a StatCan LFS-derived value. The LFS samples ~56,000 Canadian households monthly; the actual underlying sample for any given (NOC × geography) cell is thousands to tens of thousands of person-observations.

The 100-placeholder is honest about being a placeholder. The methodology paper publishes this number and notes that the reconciliation weight of 0.40 for StatCan WDS already reflects the lower confidence Orbyt assigns to this source relative to the ESDC Open Government CSV at 0.60. A future version of the connector may parse StatCan's `Annual_Wage_Notes` column (where it exists in some SEPH tables) for cell-level sample counts; the placeholder will remain until that parsing layer ships.

6.5 Confidence interval calculation

Per RFC-001's locked Estimate shape, every reconciled value carries `confidence_lower` and `confidence_upper` bounds at the specified `confidence_level` (default 0.95). The calculation is currently deterministic from the per-source sample sizes and the multi-source deviation:

- For a single-source case, the confidence interval inherits the agency's published percentile bounds where available (BLS OES publishes 10th and 90th percentile bounds directly; the Orbyt Estimate's confidence_lower / confidence_upper map to those for US base_low / base_high fields).
- For multi-source cases, the interval widens proportional to the cross-source deviation. When all sources agree closely (max deviation under 5%), the interval is narrow. When sources disagree (deviation over 25% triggers `disagreement_flag: true`), the interval widens to reflect the lower confidence.

The methodology paper notes that the confidence interval is a qualitative signal of the reconciliation engine's certainty, not a formal statistical confidence interval in the bootstrap-resample sense. A future methodology revision may introduce a true bootstrap-resample confidence interval; the version string will bump and the calculation will be documented in this paper's appendix.

---

7. Limitations and known gaps

Every compensation data product has limitations. Most products quietly omit them. Orbyt's commitment is to surface them in the methodology paper so customers can make informed product-fit decisions.

7.1 Crosswalk uncertainty

The single largest known limitation: every UK SOC ↔ US SOC mapping and every NOC 2021 ↔ UK SOC 2020 mapping is `expert-judgement`. No official government concordance exists between the United States and the United Kingdom on occupational classification. No official concordance exists between Canada and the United Kingdom either. Orbyt's mappings for these pairs are the owner's hand-curated judgment, informed by occupation titles, O*NET skill profiles, and labor-market familiarity, but they are not government-blessed.

The implication for a user: a UK estimate that maps an Orbyt slug to a specific UK SOC 2020 code rests on a judgment call. The customer should consult the lineage row's `occupation_code` field plus the mapping's `confidence` field to understand what the chain looks like. A subset of Tier 1 mappings (Canada NOC → US SOC) carry `inferred-via-NOC2016` confidence (better than pure judgment, worse than direct concordance); these are the chained mappings via the deprecated NOC 2016 intermediate.

7.2 Geographic granularity

Tier 1 coverage is metro-level, not neighborhood-level. The published geographic scope:

- **United States:** 388 metropolitan statistical areas (CBSAs) per BLS. Orbyt covers 39 of these in the initial published city taxonomy. Sub-CBSA granularity (Manhattan vs Brooklyn, Palo Alto vs Mountain View) is not in scope.
- **United Kingdom:** 12 cities (London, Manchester, Edinburgh, Bristol, Cambridge, Oxford, Leeds, Glasgow, Birmingham, Belfast, Cardiff, Reading) plus UK national plus NUTS 1-3 regional levels. Sub-city granularity (London boroughs, Manchester districts) is not in scope.
- **Canada:** 10 cities (Toronto, Montreal, Vancouver, Ottawa, Calgary, Waterloo, Edmonton, Quebec City, Halifax, Winnipeg) plus Economic Region plus province plus national. Sub-city granularity (Toronto neighborhoods, Montreal arrondissements) is not in scope.

Phase B may extend granularity where source data supports it. Today's limit is what the underlying agencies publish.

7.3 No real-time individual data

Orbyt Intelligence is not a payroll-feed product. The data sources are government statistical agencies, government regulatory filings, employer pay-transparency disclosures, and crowd-sourced submissions. None of these provide individual-payroll-record granularity.

This is a deliberate product positioning per RFC-006 §3.3. Pave and OpenComp occupy the payroll-feed segment. Orbyt occupies the developer-API-with-government-data-transparency segment. Different trade-offs apply:

- Payroll-feed products have higher individual-record fidelity but US-only or US-dominant coverage and opaque sample membership.
- Orbyt's segment has aggregated government-data fidelity, full source-attribution transparency, and ships internationally on the same methodology contract.

A customer who needs payroll-grade individual-record precision should use a payroll-feed product. A customer who needs cross-jurisdictional data with reproducible methodology should use Orbyt. The methodology paper exists to make this distinction obvious so customers can choose the right tool.

7.4 Community submissions and selection bias

For Tier C roles (no government source covers the role; sample size below the floor in available sources), Orbyt's `/api/v1/intelligence/salaries/crowd` endpoint accepts voluntary salary submissions. These contribute to the reconciliation when sample size accrues above the threshold (5 submissions per role per geography).

Selection bias in self-disclosure is real and well-documented in the broader compensation-data literature. High earners are over-represented in voluntary self-disclosure across every consumer compensation data product (Levels, Glassdoor, Payscale). Mid-career earners self-disclose more readily than entry-level or late-career; technical roles self-disclose more than non-technical; English-speaking submitters dominate the corpus.

Orbyt's `crowd` endpoint mitigates but does not eliminate this bias:

- A sample-size floor of 5 prevents tiny samples from anchoring publication
- A privacy-preserving aggregation step prevents re-identification
- A submission rate-limit prevents single users from skewing one cell

What the floor + aggregation + rate-limit do NOT do: correct for selection bias in who chooses to submit. Customers should treat Tier C estimates as biased upward in expectation, and the methodology paper recommends conservative interpretation: read Tier C medians as more likely overstating than understating the true population median by 5-15 percent, with wider error in roles where the bias is particularly steep (executive roles where only public Glassdoor- class disclosures exist; senior roles where comp-survey response rates are lowest).

7.5 Structural shifts within the survey window

Every government-data source has a survey window that ends 6-18 months before the data publishes. BLS OES May 2024 captures the labor market roughly 12-18 months before its publication. ONS ASHE October 2025 captures April 2025 jobs. ESDC's annual November release typically captures the preceding calendar year.

AI compensation has structurally shifted upward in every survey window since 2022. A role that paid $200,000 median in a 2024 survey likely pays $230,000 to $260,000 today. The survey-window-to- publication-date lag means the published Tier 1 numbers are backward-looking by design.

The methodology paper's recommendation: read Tier 1 government data as a backward-looking *floor* for AI roles in particular. Adjust upward by the role's annual growth rate (which Orbyt's `/projections` endpoint provides explicitly) when comparing against current-cycle offers. The methodology paper does not attempt a single forward- projection multiplier because the right adjustment depends on the role + the geography + the calendar quarter; the `/projections` endpoint applies the per-role growth rate to compute the forward curve.

7.6 What this product is NOT

Orbyt Intelligence is not a procurement-grade enterprise comp survey. Mercer, Aon Radford, and Willis Towers Watson occupy that segment. A procurement-grade survey buyer pays $25,000 to $100,000 per seat per year for those products. Orbyt's price point ($99 to $4,999 per month) reflects different positioning, different sales motion, and different customer expectations.

Orbyt Intelligence is not a payroll-feed black box. Pave and OpenComp occupy that segment with $50,000+ per year contracts for HR-systems-integration access to live individual-record data.

Orbyt Intelligence is not a consumer-grade entertainment tool. Levels.fyi and Glassdoor occupy that segment, with freemium pricing and opaque crowd-sourced methodology aimed at job seekers comparing themselves to their peers.

Orbyt Intelligence is a developer-API for AI compensation data priced like Stripe, accessed via REST plus MCP, with the methodology transparency this paper documents. Customers buying Orbyt to plug into their own product (a salary calculator, an offer-assessment tool, an HR analytics dashboard, an AI agent) get the methodology contract that lets them cite Orbyt in their own product without worrying that the underlying data will change shape unannounced.

---

8. Open invitations

This paper exists in part to compound. The methodology is documented not just for Orbyt's customers but for anyone building in the compensation-data space who wants a canonical reference. Three specific invitations.

8.1 The Role Taxonomy is yours

The Orbyt AI Role Taxonomy at `packages/role-taxonomy/` ships under CC BY 4.0. The license is intentionally permissive: adopt the slugs in your product, fork the taxonomy, reshape it for your domain. The attribution requirement is the only condition.

The distribution argument is straightforward. If a developer building an HR tooling product adopts Orbyt's role slugs as their canonical identifiers, every entry in that product becomes a tiny pointer back to Orbyt's compensation data. Every AI agent that uses Orbyt's taxonomy to categorize a job listing becomes a distribution channel back to Orbyt Intelligence. The taxonomy is not the paid product. The data is the paid product. Letting the taxonomy travel freely is the right trade-off.

Mapping additions and corrections are welcomed via pull request at the GitHub repository. New roles ship as MINOR version bumps; hub re-classifications ship as MAJOR with a full CHANGELOG entry per the versioning rule in §3.1.

8.2 The Methodology is open for critique

This paper is a preprint. The Orbyt repository accepts pull requests against this file directly. The repository is the canonical source; the rendered PDF and HTML are derivative artifacts.

Pull requests that point out errors, propose methodology improvements, or contribute additional citations are welcome. Errata land in the project's build log alongside the commit that addresses them. The methodology version string in this paper bumps when a substantive methodological change ships.

The intended audience for critique is broad: academic compensation researchers; HR analytics practitioners; AI agent developers using Orbyt's data; policymakers comparing national wage statistics; data journalists writing about compensation. The methodology stands or falls on the merits of the choices documented here. Surfacing disagreement publicly is a feature.

8.3 The API is priced like Stripe

Orbyt Intelligence's pricing matches the developer-API segment's expectations: four tiers from Build at $99 per month to Enterprise at $4,999 per month, monthly billing, annual discount available, no procurement gating, no minimum contract. The full pricing grid is published at /orbyt-intelligence/pricing.

A 14-day trial with Pro-tier features unlocked is included on every new subscription. The trial intentionally exposes the higher-tier features so customers can evaluate the product on its best face before committing to a permanent tier choice.

The SDK ships in TypeScript with auto-generated types from this methodology paper's underlying OpenAPI specification. The CLI ships at npmjs.com under `@orbyt/cli`. The Model Context Protocol server ships at `@orbyt/mcp` for Claude Code and other agent integrations. All three are included on every paid tier; there is no separate enterprise SKU for the SDK or CLI access.

---

9. References

References follow a hybrid arXiv-preprint style with permanent URLs to government source datasets. Each entry is grouped by category: government sources (the primary data inputs to Orbyt's pipeline), classifications (the occupation coding systems Orbyt's crosswalk operates over), prior art (compensation-data methodology papers that inform this work), and Orbyt-internal documentation (RFCs and specifications referenced inline).

9.1 Government sources

U.S. Bureau of Labor Statistics. Occupational Employment and Wage Statistics (OEWS), May 2024 release. Washington, DC: U.S. Department of Labor. https://www.bls.gov/oes/

U.S. Department of Labor, Office of Foreign Labor Certification. H-1B Labor Condition Application Disclosure Data. https://www.dol.gov/agencies/eta/foreign-labor/performance

Office for National Statistics (United Kingdom). Annual Survey of Hours and Earnings (ASHE), 2025 Provisional Release. Newport, UK: ONS. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours

HM Revenue & Customs and Office for National Statistics. PAYE Real Time Information Statistics: Monthly Seasonally Adjusted Bulletin. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/realtimeinformationstatisticsreferencetableseasonallyadjusted

UK Home Office. Skilled Worker Visa: Eligible Occupations and Codes (Appendix Skilled Occupations). https://www.gov.uk/government/publications/skilled-worker-visa-eligible-occupations/skilled-worker-visa-eligible-occupations-and-codes

Statistics Canada. Web Data Service (WDS) REST API. Labour Force Survey Table 14-10-0064-01: Employee wages by occupation. Survey of Employment, Payrolls and Hours series 14-10-0203. https://www.statcan.gc.ca/en/developers/wds

Employment and Social Development Canada. Open Government Wage Data, NOC 2021 Release. Government of Canada Open Government Portal. https://open.canada.ca/data/en/dataset/adad580f-76b0-4502-bd05-20c125de9116

European Central Bank, via Frankfurter API. Daily Foreign Exchange Reference Rates. https://api.frankfurter.dev/

9.2 Classifications

U.S. Bureau of Labor Statistics. Standard Occupational Classification (SOC) 2018 Manual. Washington, DC: U.S. Department of Labor. https://www.bls.gov/soc/2018/

Office for National Statistics (United Kingdom). Standard Occupational Classification (SOC) 2020. https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020

Statistics Canada. National Occupational Classification (NOC) 2021, Version 1.0. Government of Canada. https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1322554

Statistics Canada. NOC 2016 — SOC 2018 Concordance Table (deprecated intermediate used in Orbyt's `inferred-via-NOC2016` chained mappings).

O*NET Resource Center. Occupational Information Network database (used as a secondary reference for expert-judgement crosswalks where no official concordance exists). https://www.onetcenter.org/

9.3 Related work

Levels.fyi. Public salary database, methodology page. https://www.levels.fyi/methodology/ *Note on distinction: consumer-grade product, self-disclosed submissions, US-dominant coverage, opaque sample membership. Orbyt differs in source-attribution transparency and cross-jurisdictional scope.*

Glassdoor. Salary methodology overview. https://help.glassdoor.com/article/Glassdoor-s-Salary-Estimates/ *Note: consumer-grade product with broader role coverage than Levels.fyi but similar selection-bias structure.*

Pave. Compensation benchmarking methodology overview. https://www.pave.com/ *Note on distinction: payroll-feed product (individual-record live data from HR-systems integrations), US-only or US-dominant, opaque sample membership. Different product category from Orbyt Intelligence.*

Mercer. Pay Survey Reporting Methodology. *Note: enterprise comp-survey product. Procurement-gated. Not directly comparable to developer-API segment but referenced as the incumbent gold-standard for cross-jurisdictional compensation data.*

Hugging Face. Open Hardware Profile Standard. https://huggingface.co/ *Note on parallel: a similar canonical-structural-data publishing strategy applied to a different domain (hardware specifications). Cited as precedent for Orbyt's open-taxonomy approach.*

9.4 Orbyt-internal documentation

Bartak, J. RFC-001: Locked Response Envelope. Orbyt Intelligence Internal RFC, approved 2026-05-10. https://github.com/orbytjobs/orbyt/blob/main/docs/rfcs/RFC-001-response-envelope.md

Bartak, J. RFC-002: Query Grammar (Pagination, Filter, Sort, Expand, Scope). Orbyt Intelligence Internal RFC, approved 2026-05-10.

Bartak, J. RFC-003: Phase 2A Foundations (Reconciliation, Quarantine, Snapshots, Ingestion Pipeline). Orbyt Intelligence Internal RFC, approved 2026-05-10.

Bartak, J. RFC-004: Model Context Protocol Tool Shape. Orbyt Intelligence Internal RFC, approved 2026-05-10.

Bartak, J. RFC-006: Tier 1 International Architecture (US + UK + Canada). Orbyt Intelligence Internal RFC, drafted 2026-05-17, approved with owner-defaults 2026-05-18. https://github.com/orbytjobs/orbyt/blob/main/docs/rfcs/RFC-006-tier-1-international.md

Orbyt Intelligence. Phase 0 Verification Report (2026-05-17). Internal verification of source-URL accessibility, classification concordance availability, and Frankfurter API replacement of exchangerate.host.

Orbyt Intelligence. Owner Decisions Memo (2026-05-18). Resolution of RFC-006 open questions with defaults accepted verbatim.