How We Build China Energy Data You Can Audit

A data-quality whitepaper from WattAnalysis

Wayne Gu — WattAnalysis Working paper, November 2026

---

1. The problem: China energy data as sold today

Most China energy data reaches foreign buyers through one of three channels, and each strips out exactly the information a buyer needs to trust it.

The first channel is the single-source resale: one Chinese price assessment, one tender aggregator, one project list, translated and re-licensed. The buyer sees a number but not its provenance — which platform printed it, in what form, under which definition, and whether the number is an awarded price, a shortlist quote, or an assessor's weekly judgement. These are different things. In our own tender work we found that the procurement platforms we cover do not print awarded amounts at all: across 162 notices and 175 award records in our tender database, the count of genuine awarded-price rows is zero, structurally. What circulates in the market as "winning bid prices" for these platforms is, in large part, candidate-stage quotes relabelled. A buyer who cannot tell the two apart will misprice a market.

The second channel is the consultancy by-product. A consultancy's data exists to support its analysts' reports; it is shallow, it ages, and it carries no evidence chain, because the report — not the dataset — is the product. When the analyst leaves, the dataset dies. Nobody at the firm can show you the original document behind a figure from eighteen months ago, because keeping it was never the job.

The third channel is the scrape: high volume, no calibre discipline. Figures from different statistical cuts are summed into totals their sources never published; project lists merge "filed", "under construction" and "grid-connected" into one undifferentiated scale column; a name printed in a tender notice is matched to a

company by string similarity and quietly merged into the wrong legal entity.

Our position is that the deficit is not coverage. It is auditability. This whitepaper describes the discipline we built instead — stated as rules, then as the machine that enforces them, then as three incidents from our own build log that show the machine catching real defects, including our own.

2. Our discipline

Six rules govern every record in the six WattAnalysis databases — energy statistics, companies, projects, prices, policy measures and tenders. The databases are bilingual throughout: a Chinese master and an English mirror under identical IDs, with no Chinese text outside whitelisted verbatim-evidence fields.

Rule 1 — Verbatim evidence, archived. A value enters the database as the source prints it, or it does not enter. Original documents and page snapshots are archived in a private object store (more than 300 objects at the time of writing) and are citable if a figure is ever disputed. Where we calculate a value ourselves — customs unit values are the main case, computed as total value divided by total quantity from published customs tables — the row is labelled as our calculation and never described as a transaction price.

Rule 2 — Calibre is locked at series level. Every price series carries a five-element calibre tuple: tax basis, delivery basis, contract type, grade or specification, and system scope. The tuple is fixed for the life of the series. Changing any element does not amend the series; it starts a new one. The same separation runs through the other databases: candidate quotes and award results are different calibres and are never mixed; filed, built and grid-connected project scales are three separate fields; a statistics table's published cut and its query-channel cut are parallel observations, never summed into a single total.

Rule 3 — Conflicts are recorded in parallel, never adjudicated. When two authoritative sources print different values for the same quantity, we record both, each under its own source, and we do not pick a winner. Our nuclear series, for example, carries two published figures for the same year's generation from two official sources — 4,677 and 4,852 (in hundred-million kWh) — side by side, permanently. A buyer who needs one number can choose with full

knowledge of the other; a buyer who was never told about the conflict cannot choose at all.

Rule 4 — Every record is graded, and grade D does not ship. Each record carries a source tier and a credibility grade. Only grades A, B and C enter the sellable layer; grade D is excluded from every licensed view by default and listed as withheld. This is not a theoretical guardrail: three Shandong settlement-reference price series currently sit behind it, withheld because the publisher attribution behind them could not be resolved to our standard. They stay withheld until it is.

Rule 5 — Thin groups are suppressed, and the suppression is disclosed. No aggregate is published from fewer than three underlying observations ($n \geq 3$), and tender-price aggregates additionally require scale data. Groups that fail the test are not silently dropped and not silently published: they are entered in a suppression register — 20 groups at the time of writing — which licensees receive as part of the product. The register is, in effect, a published list of what we refuse to claim.

Rule 6 — Commercial data leaves the building only transformed. We use commercial assessment sources as inputs and cross-checks. What we sell externally is never a node-identical copy of any vendor's series: it is our cross-source composite, our index, or our model output. Per-source input values behind a composite are not licensed at any tier.

3. The machine

Rules that depend on individual care fail at scale. Ours are enforced by a small deterministic toolchain, and every merge into the repository passes through it.

The QC kernel. Each batch is checked by a QC kernel covering schema conformance, primary-key uniqueness, ID continuity, bilingual parity between the Chinese and English tables, reference integrity (every foreign key resolves), arithmetic hooks (counts and aggregates recomputed from the rows they claim to summarise), and the D-grade sellable guardrail. Exceptions are classified and must be individually accounted for in the batch record; a batch with unexplained exceptions does not merge.

The sole registrar. Formal IDs — companies, projects, sources, series — are

issued by one registrar tool and nothing else. Collectors work under provisional numbers; only the registrar mints the formal ID at merge. This sounds bureaucratic until the alternative happens: in November 2026 two of our own build streams, working in parallel from the same baseline, issued the same block of company IDs to 23 different companies. Because both streams had registered their claims, the collision was detected before merge, and the registrar re-issued the second stream's 23 entities onto a fresh block with every cross-reference rewritten and machine-verified — zero residue. In a scrape-and-merge pipeline, that collision ships silently and two companies become one.

Spot-checks and full verification. Every batch is spot-checked at no less than 15% against the original source. Where a batch's provenance is weaker than primary — a transcription, a mirror — the batch is held out of the main tables until it is verified against the issuing authority's original, in full where the document allows it.

Batch acceptance records. Every merge has a written record: what was collected, from where, what the kernel found, what was corrected, what remains open. The three cases below are taken from those records.

Case 1: The notices that delete themselves

State Grid's procurement platform publishes candidate (shortlist) notices in two forms. Attachment-form notices can be downloaded later. Embedded-form notices cannot: when the 1–3 day publicity window closes, the platform permanently replaces the notice body with a stub reading, in effect, "publicity completed." There is no recovery channel; the platform's bulk-download endpoint for an expired notice returns an empty package.

We discovered this by measuring it. In a full-census mining run (batch QC #85), of 320 embedded-form candidate notices examined, 15 still carried their content — 4.7%. The other 305 were stubs, including 82 whose titles marked them as materials-procurement batches with real scale. Roughly a quarter of the embedded census was structurally unrecoverable, and no amount of later effort can rebuild it: the originals no longer exist anywhere.

The operational consequence is now a standing rule of the house: candidate notices are captured daily, inside the publicity window, and the originals are archived the same day. (A second platform we cover, CHN Energy's, does not

destroy its notices — its January 2026 candidate pages were still full-text in November — so it is swept monthly instead. The capture cadence follows the evidence about each platform, not a uniform schedule.) Every day this archive grows, it becomes less reproducible by anyone who starts later. That is not a marketing claim; it is a property of the source.

Case 2: Verification that corrects the record — including ours

A project-expansion batch drew 294 rows from non-primary carriers: provincial lists transcribed or mirrored by third parties. The rows were held at a lower tier until each province's list could be checked against the issuing authority's original.

For Shanxi's 63 projects, the parent document was located on the provincial energy bureau's own site, and its attachment PDF proved byte-identical — matching MD5 — to the copy we had collected. All 63 rows were upgraded to primary tier on that evidence. For Hebei's 21 projects, the original was a stamped PDF on the Qinhuangdao development and reform commission's site, with no text layer; verification proceeded by transcription and programmatic comparison of name, owner and scale. Eighteen rows matched verbatim. Three did not: our collected names contained transcription errors — a wrong character in one project name, a missing slash in a "200MW/800MWh" designation in two others. The original governs; all three were corrected, and each correction is itemised in the batch record. The same document also contained an inconsistency of its own — one project's name states 100MW/400MWh while its capacity column prints 200MW/800MWh. We record the capacity column as printed, keep the name as printed, and note the discrepancy on the row. We do not resolve it, because it is not ours to resolve.

Two provinces — Guangxi (203 rows) and Anhui (7) — could not be verified: the official originals exist but were unreachable from every network route available to us. Those rows remain at the lower tier, unmerged into the headline counts in this paper, with the retry conditions written into the record.

Case 3: The defect that passed every shape test

A merge-assembly defect in our own company register shifted the last five fields of four rows one column to the left: an empty parent-company field had been dropped during assembly, and the remaining values slid into its place. The

rows were the correct width. The schema check passed. Only the semantics were wrong — a batch code sat where a parent company ID belonged, a count column held prose, the provenance note was empty.

The kernel caught it, because reference integrity is a semantic check: a parent-company ID that resolves to nothing is an exception regardless of how well-formed the row looks. The repair was then done the way the discipline requires — per column, per row, from independent evidence: the original registration sheet (which has no parent field by design), neighbouring rows of the same form, and a manual recount of each company's actual references in the tender tables (each of the four is referenced exactly once). The correction cleared exactly the 12 kernel exceptions the defect had caused — eight dangling-parent flags and four bilingual count mismatches — and introduced zero new ones. The batch record states plainly what the fix does not claim: restoring the empty parent field restores the register's design; it is not an assertion that these companies have no parent.

4. What this yields

Cross-source composites. Our flagship price form prices the same node from several independent assessment sources and publishes only the composite: the median, a dispersion band, and the source count. Two series are live. WA-PRC-28 (battery-grade lithium carbonate, daily) carries 5 points; on 20 November 2026 its three inputs printed 91,450, 91,600 and 91,500 CNY per tonne, for a composite of 91,500 at 0.16% dispersion — and the revision history of that point is itself recorded, because one source's evening-session value replaced its morning substitute. WA-PRC-29 (280 Ah storage cells, weekly) carries 7 points back to August 2025. Where our checks found a systematic gap between two sources — one assessing on Fridays, the other publishing weekly averages, a level difference of 0.007–0.015 CNY/Wh — we registered it as a calibre difference, disclosed it, and did not "correct" either source.

Derivative indices with published methods. Two index families are built only from the layers above, each preceded by a written method note. The customs unit-value index (UVI) publishes 64 quarterly points back to 2020Q2; one candidate component — solar cells — was excluded from the index because only 2 of its 21 quarters are complete, an upstream data-density problem that no index construction should paper over. The tender reference index (TCQ)

published its first formal point for 2026Q3: seven first-ranked candidate quotes, a capacity-weighted 0.6949 CNY/Wh, the full sample list disclosed. It is named, in every context, "first-ranked candidate quotes" — never "awarded prices" — because that is what the underlying records are.

Cost models, labelled as models. Quarterly frozen runs — a PV full-chain cost curve (46 points), a storage BOM/LCOE model (28 points), a green-hydrogen LCOH model (19 points) — carry the label "model output, not an observation" on every row, with assumption sets available to licensees.

A falsifiable prediction register. The model library's forecasts are entered into a standing register before their outcome windows open — 40 predictions are currently on record: 18 pre-registered with the first model runs of 17–18 November 2026, and 22 sealed on 24 November 2026 — each dated at registration, scored against the printed outcome when the window closes, and kept on the record whether it hits or misses. A forecast that cannot be wrong is marketing; ours are written so that they can be.

5. Honest limits

A quality discipline that only reports successes is marketing. Our current gaps, stated as the roadmap they define:

- **Tenders: 162 notices against a 600 target.** Two of three planned platforms are in routine capture; the third is not yet in the monthly cycle. The binding constraint on tender-price aggregates is scale data — most notices do not print quantities — and one index family (TCQ-SYS) consequently has no points yet.
- **Companies: 1,131 entities against a 1,200 target.** Internal evidence is exhausted — 98% or more of the parties already appearing in our tender and project tables are registered — so the remaining gap closes only through new-source collection, at collection speed.
- **Composite depth is source-bound, not effort-bound.** Twelve formal composite points exist against a target of forty. The constraint is counterpart availability: our principal weekly source for cell prices has no regular public counterpart — of 43 weekly issues, 33 weeks have no second-source value to pair — so depth accumulates at the speed of the weekly routine. A weekly lithium-carbonate composite was evaluated

and honestly recorded as zero defensible points (the single candidate pair showed 14.4% dispersion and a timing mismatch); it will be re-evaluated once the routine has accumulated at least eight same-week overlaps. A polysilicon composite is in pilot with two trial points and takes a formal series number only after four weeks of routine capture.

- **Projects: target met, verification incomplete.** 858 projects across 11 of 12 target provinces. Guangxi's 203 rows and Anhui's 7 remain unverified at a lower tier (see Case 2), and 68 rows carry placeholder county codes; placeholder rows are barred from aggregation until the codes are verified.
- **Statistics:** the core frames are closed and current to September 2026; depth now grows by the monthly cycle, one month at a time.

Each of these is a dated work item in our build plan, not an apology. The scorecard we use to grade ourselves — ten dimensions, re-scored after every build wave, with the shortfalls written next to the scores — is the same document our catalogue figures are drawn from.

6. How to work with us

Two products sit on top of this evidence base.

The Analyst subscription (US$400/year, public) is the reading product: our published analysis and the public single-series data surface.

The Compliance Desk (from US$1,200/year, public anchor) is the judgement layer for a client's own exposure: measure-level tracking of the trade and energy measures that touch their products — 59 structured measures across China, the US, the EU, India and Canada, each with market, direction, product, HS codes, rate text as printed, effective window and status — with dated-risk alerts ahead of effect. Expiries and conditional revivals are modelled as states with dates and re-checked before they bite: at the time of writing, an export-control suspension lapses on 27 November 2026 and the prior announcement revives the following day, and both states are rows in the database, not footnotes in a newsletter.

The Data Licence grants use of the six databases themselves, priced by use right rather than by the row, in three tiers — Individual, Team (with key-gated

API access and entity resolution), and Data Redistribution (with the suppression registers and withheld-series lists included, so a licensee's product inherits the same honesty constraints). Licence tiers are not publicly priced; they are quoted per proposal, because the right being sold differs by buyer. What never changes by tier: composites are licensed, their per-source inputs are not; model outputs keep their label downstream; grade D stays out.

Founding terms: the first three Compliance Desk engagements signed before 31 March 2027 lock their year-one rate for two renewal years and receive the Data Licence (Individual tier) at no additional charge for the first year.

WattAnalysis — contact@wattanalysis.com. Data catalogue and sample pack available on request. Figures in this draft: repository census at head `1c3f8a2`, 2026-11-21; build batches cited as QC #85, QC #95-A2 and QC #95-B (fix640 record).