Status: Internal only. The rendered PDF is at `docs/reliability/DELIVERABLE_QUALITY_2026_Q3.pdf` (also mirrored to `frontend/public/docs/reliability/DELIVERABLE_QUALITY_2026_Q3.pdf`). Numbers are illustrative until `reliability_events` Stage O (Supersedes / contradiction) telemetry is populating `pipeline_triples`. This file is the editorial version; the production query is at `agents/src/services/reliability_events.py` Stage `O` payload, queried over `pipeline_execution_events`.
Plan reference: `docs/strategy/EXECUTION_AND_MARKETING_PLAN_2026_08_17.md` §3 M-4.
Why this is the play: None of the 56 competitors in the Aug 18 research file publicly ship accuracy metrics. KPMG publishes agent count. Concourse publishes client outcomes. Nobody publishes "the bridge reproduced within ±X% on a held-out test fixture." We can, because `pipeline_triples` (the E-2 table) carries the evidence_quote of the source cell alongside every produced cell.
Cadence: Updated quarterly. First public version is gated on 2 quarters of consistent data.

Defect-rate table (Jul 2026 — illustrative, team to fill)

Pipeline	Mean defect rate	Sample size	Period	Method
QoE Adjusted-EBITDA bridge	0.0%	31	Jul	±0.5% on golden-output benchmark (`tests/fixtures/qoe_july.xlsx`, golden run `4ee56aa0`)
Multi-source reconciliation	2.1%	14	Jul	per-figure row-by-row vs golden (Stage O contradiction_kind = `value_diff`)
Inventory roll-forward	4.7%	8	Jul	closing-units ≡ opening + net change; per-cell rule verification
Pipeline-built reporting	(not yet measured)	0	—	methodology draft (per Aug 19 PR `23155cb6`, cleansing-chain routing must be excluded for multi-input pipelines)

Methodology footnotes:

1 *QoE Adjusted-EBITDA bridge.* The ±0.5% reproduction rule means: for a deliverable that should produce 1500/1075/1389/1536 (the golden run values), every cell must reproduce within ±0.5% of that figure across re-runs. A defect is a cell that fails this rule on a re-run with the same input. After Aug 19 PR

`ef9b4ff1` (empty runs surface as `NEEDS_REVIEW` not `COMPLETED`), an empty binding-result is *not* a defect — it's a binding failure, recorded separately as `NEEDS_REVIEW` rather than counted in this table.

2. *Multi-source reconciliation.* Defect = cell whose value differs from the prior golden, where the prior golden and the new run disagree on the same `(source_doc, source_column)` row (Stage O fires). The four contradiction_kinds tracked: `value_diff`, `evidence_diff`, `unit_diff`, `sign_flip`. See `agents/src/services/reliability_events.py` Stage `O` payload.
3. *Inventory roll-forward.* This is the only one of the four where the defect definition is structural, not statistical: the closing units must equal opening + net change to the unit. A defect is a row where that invariant fails.
4. *Pipeline-built reporting.* Methodology only; we don't have a benchmark yet. The next step is a frozen reporting deliverable (likely a Trial Balance w/ variance to prior period) and a golden-run fixture.

Production data path (for the team that will fill this in)

Important: The Stage O telemetry only becomes load-bearing once E-2 (per-cell provenance sidecar) ships, because E-2 is what writes the `pipeline_triples` rows that Stage O compares. Until E-2 PR-1 lands, this table is filled by hand from `reliability_events` log inspection. See plan Appendix D.1 for E-2 PR boundaries.

Roll-out plan (per plan §3 M-4)

1. **Internally first.** Posted to `finadvantage` Slack `#reliability` + a private Notion URL.
2. **Reviewed quarterly.** CEO + Reliability Lead sign-off on each refresh.
3. **Public.** Only after 2 quarters of consistent data, and only with the "Method" column shown verbatim (no marketing rewrite). The "Method" column is what makes this not just marketing — it shows we did the comparison. Plan §3 M-4: *"Numbers are illustrative; the team fills in. The structure is the play."*

What this play explicitly does NOT do

- Does NOT publish per-customer defect rates. Aggregate only.
- Does NOT publish accuracy metrics on deliverables where we have no golden-output benchmark (e.g. free-form pipeline-built reporting).
- Does NOT publish numbers we can't back up with `pipeline_triples` rows. The Aug 19 research shows KPMG and FloQast publish broad claims ("auditable by design"); we publish *counted* claims only.
- Does NOT contradict Move M-8 (outcome-based pricing). The 0.0% QoE defect rate is *exactly* the basis for the ±0.5% reproduction rule that M-8's outcome-pricing PDF promises. The two plays anchor each other.