# Evidence Assessment Case Study

## AI-Powered Candidate Screening System

**Client:** [REDACTED] — Series B HR-tech, Berlin **System:** Automated resume screening and ranking for enterprise recruiters **EU AI Act Classification:** High-risk (Annex III, Section 4 — Employment) **Assessment Duration:** 5 business days

---

## Before vs. After

| | Before Our Assessment | After Our Assessment |
|---|---|---|
| **CTO's estimate** | "A few weeks to write the docs" | 120–160 hours of actual remediation work |
| **Known gaps** | 0 — assumed they were "mostly ready" | 4 gaps found, including 2 critical |
| **Bias testing** | "We don't discriminate" (no evidence) | Critical gap: zero fairness evaluation across protected categories |
| **Evidence sources** | "It's all in Confluence somewhere" | Scattered across 7 systems (GitHub, MLflow, Confluence, Datadog, Terraform, Swagger, internal admin) |
| **Compliance risk** | Unknown | 2 issues that would likely fail a conformity assessment |
| **Remediation plan** | None | Prioritized roadmap with owners, effort estimates, and deadlines |

---

## The Situation

The client's AI system processes 50,000+ applications per month across 12 enterprise customers. With the August 2, 2026 deadline approaching, leadership needed to understand their compliance readiness — but had no internal expertise on Annex IV requirements.

Their CTO estimated it would take "a few weeks" to prepare documentation. After our assessment, the actual evidence landscape was significantly more complex than expected.

---

## Evidence Gap Assessment Results

### Summary Dashboard

| Metric | Result |
|---|---|
| Evidence areas assessed | 9 of 9 (Annex IV) |
| Fully documented | 3 of 9 |
| Partially documented | 4 of 9 |
| Critical gaps | 2 of 9 |

| | |
|---|---|
| Estimated remediation effort | 120–160 engineering hours |
| Estimated time saved by assessment | 80+ hours |

---

## Detailed Findings by Annex IV Section

### Section 1 — General Description of the AI System

**Status: COMPLETE**

| Evidence Required | Source Found | Location |
|---|---|---|
| Intended purpose & scope | Product documentation | Confluence: `/product/ai-screening-overview` |
| System interaction description | API documentation | Swagger: `api.screening.internal/docs` |
| Hardware/software requirements | Infrastructure specs | Terraform configs: `infra/prod/` |
| Version history | Release notes | GitHub Releases: 47 tagged versions |

**Assessment:** Well-documented. Minor gaps in version-to-version change descriptions. Estimated fix: 4 hours.

---

### Section 2 — Development Process & Data

**Status: PARTIAL — 2 sub-gaps identified**

| Evidence Required | Source Found | Location | Status |
|---|---|---|---|
| Design specifications | Architecture docs | Confluence: `/eng/ml-architecture` | Found |
| Training methodology | Experiment tracking | MLflow: experiments #201–#847 | Found |
| Training data description | Data documentation | **No structured data sheet** | GAP |
| Data governance measures | Data pipeline code | `data-pipeline/preprocessing.py` | Partial |
| Bias examination & mitigation | Bias testing results | **Not found in any system** | CRITICAL GAP |
| Validation & testing approach | Test suite | `tests/model/` — unit tests only | Partial |

**Critical Finding:** No structured bias testing framework exists. The model processes candidate data across protected categories (age, gender, nationality) but there is no documented evidence of fairness evaluation. This is the highest-priority remediation item.

**Assessment:** Training methodology is well-tracked in MLflow, but the data governance documentation and bias testing evidence are either missing or not systematically recorded. Estimated fix: 40–60 hours (bias framework: 30h, data sheets: 10h, validation expansion: 10–20h).

---

### Section 3 — Monitoring, Control & Human Oversight

**Status: PARTIAL**

| Evidence Required | Source Found | Location | Status |
|---|---|---|---|
| Human oversight measures | Override mechanism | `src/api/manual-review.ts` | Found |

| | | | |
|---|---|---|---|
| Logging capabilities | Application logs | Datadog: `screening-service` | Found |
| Override/intervention capability | Admin dashboard | `admin.screening.internal` | Found |
| Deployer instructions | Customer documentation | **Incomplete — no AI-specific guidance** | GAP |

**Assessment:** Technical controls exist but deployer-facing documentation lacks AI-specific guidance on human oversight responsibilities. Estimated fix: 16 hours.

---

**Section 4 — Accuracy, Robustness & Cybersecurity**

**Status: PARTIAL**

| Evidence Required | Source Found | Location | Status |
|---|---|---|---|
| Accuracy metrics | Model evaluation | MLflow: F1=0.87, AUC=0.92 | Found |
| Robustness testing | Adversarial tests | **Not found** | GAP |
| Performance across subgroups | Disaggregated metrics | **Not found** | CRITICAL GAP |
| Cybersecurity measures | Security documentation | SOC 2 Type I report | Found |
| Known limitations | Error analysis | `docs/known-issues.md` — outdated | Partial |

**Critical Finding:** Model accuracy metrics exist at aggregate level but no disaggregated performance analysis across demographic subgroups. Combined with the Section 2 bias gap, this represents the client's largest compliance risk.

**Assessment:** Cybersecurity posture is strong (existing SOC 2). Accuracy tracking exists but lacks the granularity regulators expect. Estimated fix: 20–30 hours.

---

**Sections 5–9 — Risk Management, Lifecycle, Standards, Declaration, Post-Market**

**Status: COMPLETE (Sections 5, 7) / PARTIAL (Sections 6, 8, 9)**

Key findings:

- **Risk management system** (Section 5): Documented in `docs/risk-assessment.md` with quarterly review process. Complete.
- **Lifecycle changes** (Section 6): Git history provides full traceability, but no structured change impact assessment process for model updates. 8 hours to remediate.
- **Applied standards** (Section 7): SOC 2 and GDPR compliance documented. Complete.
- **EU Declaration of Conformity** (Section 8): Not yet drafted. 4 hours with legal counsel.
- **Post-market monitoring** (Section 9): Datadog monitors system health but no structured process for monitoring AI-specific performance degradation. 12 hours to remediate.

---

## Remediation Roadmap

### Priority 1 — Critical (Blocks Compliance)

| Action | Owner | Effort | Deadline |
|---|---|---|---|
| Implement bias testing framework | ML Team | 30h | Week 1–3 |
| Create disaggregated performance metrics | ML Team | 15h | Week 2–4 |

| | | | |
|---|---|---|---|
| Document training data governance | Data Team | 10h | Week 1–2 |

### Priority 2 — High (Required for Annex IV)

| Action | Owner | Effort | Deadline |
|---|---|---|---|
| Expand validation test suite | ML Team | 15h | Week 3–5 |
| Create deployer AI oversight guide | Product Team | 16h | Week 2–4 |
| Establish model change impact process | Eng Lead | 8h | Week 3–4 |

### Priority 3 — Standard (Completion Items)

| Action | Owner | Effort | Deadline |
|---|---|---|---|
| Set up AI performance monitoring | Platform Team | 12h | Week 4–6 |
| Draft EU Declaration of Conformity | Legal + Eng | 4h | Week 5–6 |
| Update version change descriptions | Eng Lead | 4h | Week 1–2 |

**Total estimated remediation: 120–160 hours across 6 weeks**

---

## Outcome

The client's CTO initially estimated "a few weeks" to handle Annex IV documentation internally. Our assessment revealed that the documentation itself was only 20% of the challenge — the remaining 80% was identifying, extracting, and structuring evidence that was scattered across 7 different systems (GitHub, MLflow, Confluence, Datadog, Terraform, Swagger, internal admin tools).

The two critical gaps identified (bias testing and disaggregated metrics) would likely have been discovered during a conformity assessment — at a point where remediation timelines would be far more constrained and costly.

---

*This case study is based on a composite of real assessment patterns. Client details have been anonymized. EU AI Ready prepares technical compliance evidence — we do not provide legal advice.*