

Das Datenqualitäts-Playbook: Agentic AI mit Freigabe-Gates

Warum KI-Projekte an schmutzigen Daten sterben, wie agentische Bereinigung ohne Kontrollverlust funktioniert — und die Kennzahlen, die es beweisen.

Die Ein-Absatz-Version: Gartner prognostiziert, dass Organisationen bis 2026 **60% der KI-Projekte** abbrechen, die nicht durch KI-taugliche Daten gestützt sind. Dubletten, Formatabweichungen und fehlende Identifikatoren sind kein IT-Ärgernis — sie sind der Grund, warum Automatisierungsinitiativen stecken bleiben. Agentische KI kann Stammdaten semantisch bereinigen (sie erkennt, dass «Müller AG», «Mueller AG» und «Müller AG, Grenchen» ein Kunde sind), aber erst das menschliche Freigabe-Gate macht das Ergebnis vertrauenswürdig: **Die KI schlägt vor, Menschen geben frei.**

Die vier Fehlermuster schmutziger Stammdaten

- **Dubletten, die keine Regel-Engine findet** — Namensvarianten, Umlaute, Abkürzungen, umgezogene Adressen
- **Formatabweichungen** — Telefonnummern, Daten und IDs nach jedem Systemwechsel anders geschrieben
- **Fehlende Identifikatoren** — Datensätze, die sich über ERP, CRM und Tabellen nicht verknüpfen lassen
- **Widersprechende Systeme** — ERP, CRM und Excel «wissen» je eine andere Adresse oder einen anderen Preis

Warum Agenten — und warum Kontextfenster zählen

Klassische Datenqualitäts-Tools führen handgeschriebene Regeln aus; Agenten lesen Daten wie eine sorgfältige Mitarbeiterin. Aktuelle Claude-Modelle halten 200K–1M Tokens Kontext (rund 300–2'500 Seiten oder Zehntausende Datensätze) — ein Agent kann Ihr Schema, Ihre Geschäftsregeln und grosse Stichproben gleichzeitig betrachten. Publierte Forschung (NVIDIA RULER, Chromas Context-Rot-Studie 2025) zeigt aber: Der effektive Kontext ist kleiner als beworben. Ein seriöses Setup kombiniert deshalb langen Kontext fürs Reasoning mit direkten MCP-Datenbankabfragen und Batch-Läufen. Nie eine Datenbank in einen Prompt kopieren.

Das Betriebsmodell in vier Schritten

1. **Audit:** Agenten profilieren ERP, CRM und Dateien; bewerteter Bericht über Dubletten, Lücken, Inkonsistenzen — je mit Geschäftswirkung.
2. **Bereinigung mit Freigabe-Gates:** Korrekturen und Zusammenführungen in Batches vorgeschlagen; ein Mensch gibt frei; ohne Freigabe wird nichts zurückgeschrieben.
3. **Validierung an der Quelle:** Neue Datensätze werden bei der Erfassung geprüft — Plausibilität, Vollständigkeit, systemübergreifende Konsistenz.
4. **Monitoring:** Ein Daten-Gesundheits-Dashboard und Anomalie-Alerts, damit Qualität nicht mehr still degradiert.

Die Kennzahlen, die es überprüfbar machen

- Dublettenrate** — vor dem ersten Bereinigungs-Batch gemessen, nach jedem verfolgt
- Fehlerrate auf kritischen Feldern** (Identifikatoren, Adressen, Zahlungskonditionen)
- Stunden manueller Bereinigung pro Monat** — die Zahl, die Ihr Team spürt
- Datensätze, die an der Eingangsvalidierung scheitern** — der Frühindikator für künftiges Chaos

Die Baseline wird im Audit definiert; danach sehen Sie die Zahlen nach jedem Batch. Bewegen sie sich nicht, sehen Sie auch das — genau das ist der Punkt.

Der Marktkontext (zum Vergleichen)

Gartner hat die Kategorie 2024 in «Augmented Data Quality Solutions» umbenannt; der Magic Quadrant 2025 nennt Ataccama, Informatica und Qlik als Leaders. Das sind exzellente Enterprise-Plattformen mit Enterprise-Zeitplänen und -Budgets. Die Alternative in KMU-Grösse: Claude + MCP gegen die Systeme, die Sie bereits betreiben (bexio, ABACUS, Microsoft 365) — Wochen statt Quartale, Fixpreise, Human-in-the-Loop by Design.

Dieser Leitfaden ist die kompakte Fassung unseres publizierten Fachartikels — Quellen und Vertiefung: eflury.com/de/services/datenqualitaet/. Stand: Juli 2026. Keine Rechtsberatung.

Fragen zur Umsetzung in Ihrem Betrieb?

Kostenloses 30-Minuten-Gespräch: eflury.com/de/kontakt/ · me@eflury.com · +41 79 910 77 87