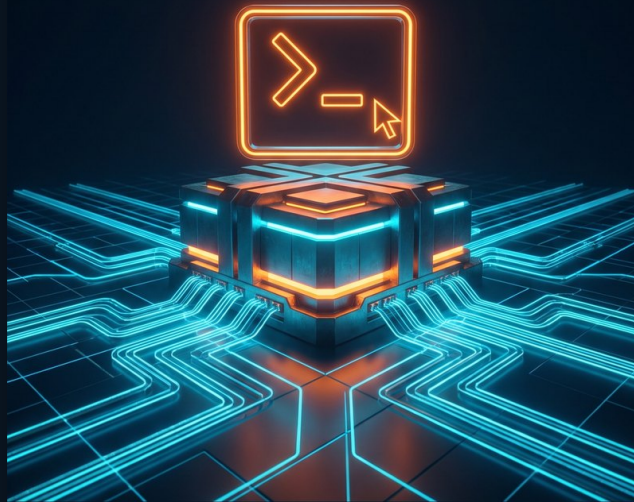


MLOps · LLMOps · Data Engineering

Betrieb statt Demo



Thorsten Behder

Die Go-Live-Checkliste für KI-Systeme

Begleitpaket zum Buch „Betrieb statt Demo“

von Thorsten Behder

Die Vorlagen aus dem Buch – Data-Contract-Skeleton, SLO-Sheet, Rollout-Plan und Eval-Set – als Go-Live-Checkliste, bevor ein KI-Produkt produktiv geht.

CHECKLISTE

Diese sieben Arbeitsblätter stammen wörtlich aus Anhang C des Buchs. Die Platzhalter in eckigen Klammern füllen Sie mit den Angaben Ihres eigenen Systems.

Sieben Vorlagen für den KI-Betrieb

Aus dem Buch — Anhang C: Vorlagen und Arbeitsblätter

1. Data-Contract-Skeleton

```
# datacontract.yaml (Skeleton nach Data Contract Specification)
info:
  title: ""          # sprechender Name der Datenquelle
  owner: ""          # verantwortliches Team, kein Einzelname
  contact: ""        # erreichbarer Kanal, kein Sammelpostfach
models:
  <tabellenname>:    # ein Modell je Tabelle oder Topic
    fields:
      <feld_id>:
        type: string  # string | decimal | integer | date | ...
        required: true
        unique: true  # nur für Schlüsselfelder
      <feld_kritisch>:
        type: decimal
        required: true
        description: "" # Einheit, brutto/netto, Bezugszeitpunkt!
        quality:
          - type: range # Regel auf Werteebene, faengt stille
            min: 0      # Semantikaenderungen (Euro -> Cent)
            max: 0      # realistische Obergrenze eintragen
          # ... weitere Felder nach Kritikalitaet, nicht Vollstaendigkeit
servicelevels:
  freshness:
    threshold: PT6H   # ISO-8601-Dauer; muss zum Nutzungszweck passen
```

2. SLO-Sheet für ein KI-Produkt

Dienst: [Name und Endpunkt oder Schnittstelle]
SLI: [Was wird gemessen, wo und wie genau?]
Zielwert: [Zahl mit Vergleichsoperator]
Messfenster: [rollierend X Tage / Kalendermonat]
Error Budget: [was der Zielwert an Fehlern uebrig laesst]
Owner: [Team, das den Wert beeinflussen kann]
Eskalation: [was bei 50 % Verbrauch passiert, was bei 100 %]

3. Rollout-Plan

ROLLOUT-PLAN

Aenderung: [was genau wird ausgerollt, Version alt -> neu]

Risikoklasse: [niedrig / mittel / hoch, mit Ein-Satz-Begründung]

Verfahren: [Shadow / Canary / Blue-Green, ggf. verkettet]

Stufen: [z. B. Shadow 2 Wochen -> 5 % für 1 Woche
-> 25 % für 1 Woche -> 100 %]

Zuteilung: [randomisiert pro Anfrage/Vorgang; Repräsentativitäts-
Check der Canary-Verteilung gegen Gesamt-Traffic]

Vergleichsmetriken:

- [technisch: Latenz, Fehlerrate]
- [fachlich: Score-Verteilung, Eval-Ergebnis, Ablehnungsquote]

Abbruchkriterien:

- [Guardrail-Verletzung X -> automatischer Stopp]
- [Beobachtung Y -> Pause und Analyse]
- [wer den Abbruch aussprechen darf]

Rollback-Weg: [wie zurueckgeschaltet wird, wie lange es dauert,
wann zuletzt geprobt; Modell UND Prompt UND Config]

Beobachter: [wer die Metriken wann ansieht, namentlich als Rolle]

4. Eval-Set-Vorlage

```
# eval_case.yaml (ein Fall; das Set ist eine Sammlung solcher Faelle)
id: EV-0042      # stabil, wird nie wiederverwendet
input: ""        # die Eingabe, so realistisch wie möglich
soll: ""         # Soll-Kriterium: was muss die Antwort
                # enthalten, was darf sie nicht enthalten
pruefart: regel  # regel | judge | mensch
                # regel: maschinell pruefbar (Klausel enthalten,
                # Quellenangabe vorhanden, Format korrekt)
                # judge: LLM-Judge mit Bewertungsanweisung
                # mensch: Stichprobe durch Fachexperten
gewicht: 1       # kritische Faelle hoeher gewichten
quelle: ""       # Herkunft: echter Fall, Ticket, Beschwerde
stand: 2026-07   # wann aufgenommen oder zuletzt geprueft
```

5. Kostenmodell-Rechenblatt

KOSTENMODELL [System, Stand JJJJ-MM]

Anfragen pro Tag: [N]

Vorgaenge pro Tag: [V] (ein Vorgang kann mehrere
Anfragen ausloesen)

Kettenstufen (eine Zeile je Modellaufruf in der Kette):

Stufe 1: [Zweck] Input stabil: [Ts] Tok Input variabel: [Tv] Tok
Output: [To] Tok Modell: [M1]

Stufe 2: [Zweck] ...

Agenten: mittlere Schrittzahl ansetzen, nicht die beste

Cache-Trefferquote (stabiles Praefix): [H] % (aus der API-Antwort,
als Betriebsmetrik mit Alarm)

Preise (Platzhalter, aktuelle Preisliste einsetzen):

P_in = [] EUR / 1 Mio. Input-Tokens

P_out = [] EUR / 1 Mio. Output-Tokens

P_cache_read = [] EUR / 1 Mio. Tokens (typisch Bruchteil von P_in)

Kosten pro Anfrage =

Summe über Stufen: $(Ts * (1-H) * P_{in} + Ts * H * P_{cache_read}$
 $+ Tv * P_{in} + To * P_{out}) / 1.000.000$

Kosten pro Geschaeftsvorgang = Kosten pro Anfrage * (N / V)

6. AI-Act-Einordnungsbogen

AI-ACT-EINORDNUNGSBOGEN

System: [Name, Version, Verweis auf Systemdoku]

Zweckbestimmung: [ein Satz, woertlich wie kommuniziert
und tatsaechlich betrieben]

Betroffene Personen: [wer, Entscheidungsfolgen, wer sieht das
Ergebnis und was tut er damit]

Verbotsprüfung (Art. 5): [einschlaegig ja/nein, Begrueendung]

Annex-III-Pruefung: [jeder ernsthaft in Betracht kommende
Tatbestand einzeln: Fundstelle, Merkmale,
Begrueendung der (Nicht-)Erfuellung]

Transparenzpflichten (Art. 50): [welche greifen, wie umgesetzt]

GPAI-Bezug: [genutzte Basismodelle, eigene Rolle:
Betreiber/Kunde, relevante Anbieterzusagen]

Dokumentationsbedarf: [was daraus folgt: Art.-6-Abs.-3-Assessment,
Registrierung, technische Doku]

Kipp-Trigger für Neubewertung:

- [Aenderung der Zweckbestimmung]
- [Automatisierung einer bisher menschlichen Entscheidung]
- [neue Nutzergruppe oder Produktlinie]
- [Aenderung der Rechtslage]

Pruefdatum / Pruefer: [JJJJ-MM-TT, Rolle]

Jurist-Review: [ja/nein, wer, Datum der Gegenzeichnung]

7. Postmortem-Vorlage

POSTMORTEM: [kurzer, sachlicher Titel des Vorfalls]

Schweregrad: [z. B. SEV-1 bis SEV-3 nach eigener Skala]

Status: [Entwurf / reviewt / abgeschlossen]

ZEITLEISTE (minutengenau, nur Rollen)

[JJJJ-MM-TT hh:mm] Beginn des Vorfalls (auch wenn ihn da noch niemand bemerkte)

[hh:mm] ... [Ereignis, wer oder was es bemerkte]

[hh:mm] Entdeckung -> Zeit Beginn bis Entdeckung: [X]

[hh:mm] Behebung -> Zeit Entdeckung bis Behebung: [Y]

AUSWIRKUNG

[betroffene Nutzer/Vorgaenge, Dauer, geschaeztzte Kosten, falsche Entscheidungen, die auf dem System beruhten]

BEITRAGENDE FAKTOREN (drei bis sechs, kein Alleintaeter)

1. [Bedingung, die für sich harmlos gewesen wäre]

2. ...

WAS GUT LIEF

[was Entdeckung oder Behebung beschleunigt hat; ehrlich]

MASSNAHMEN (was, Owner, Termin; sonst keine Massnahme)

1. [was gebaut/geaendert wird] Owner: [Rolle] bis: [Datum]

2. ...

VERTEILER

[wer das Dokument liest: Team, Nachbarteams, Leitung;
wo es abgelegt und auffindbar ist]

Die ausgefüllten Beispiele zu diesen Vorlagen und die zugehörigen Kapitel stehen im Buch.

Dieses Paket gehört zum Buch „Betrieb statt Demo“

<https://www.amazon.de/dp/BOHBN1ZDPR>

tbai.cloud/buch-betrieb-statt-demo.html

Autor und inhaltlich verantwortlich

Thorsten Behder

Weilheimer Str. 21A

81373 München

thorsten@behder.de



Gestaltung und Layout: tbai.cloud