

Warum das Modell im Betrieb beobachtet wird

Ein Sprachmodell meldet sich nicht, wenn es schlechter wird. Es antwortet einfach weiter.

Im Test kennen Sie die richtige Antwort. Im Betrieb kennen Sie sie nicht.

Deshalb misst Monitoring Auffälligkeiten statt Richtigkeit.

Was ein Monitoringlayer ist

Eine Schicht zwischen Ihrer Anwendung und dem Modell, die jeden Aufruf mitschreibt und laufend auswertet: was hineinging, was zurückkam, wie lange es dauerte, was es kostete und ob das Ergebnis die erwartete Form hat. Sie greift nicht ein. Sie sorgt dafür, dass eine Veränderung sichtbar wird.

Fünf Gründe, warum es ohne nicht geht

1 Der Test deckt nur ab, was in ihm vorkommt.

Der Betrieb bringt Dokumente, Formulierungen und Sonderfälle, die beim Aufbau niemand kannte.

2 Fehler sind still.

Es stürzt nichts ab. Ein falscher Wert sieht aus wie ein richtiger und wird weiterverarbeitet.

3 Die Eingaben verschieben sich schleichend.

Neue Layouts, andere Sprachen, ein neuer Lieferant. Jede Änderung für sich ist unauffällig.

4 Das Modell wird ausgetauscht.

Anbieter aktualisieren Fassungen und stellen ältere ab. Das wirkt sofort im Produktivpfad.

5 Kosten und Laufzeit hängen an der Eingabe.

Ein längeres Dokument kostet mehr und dauert länger. Ohne Messung fällt es auf der Rechnung auf.

Ein Modell ohne Beobachtung ist ein Lieferant ohne Wareneingangskontrolle.

Sechs Signalgruppen ohne Sollwert

Im Betrieb ist das richtige Ergebnis unbekannt. Diese Signale zeigen trotzdem, wenn etwas kippt.

1 Technische Signale

Laufzeit je Aufruf, fehlgeschlagene Aufrufe, Zeitüberschreitungen, Wiederholungen.

Wirkung: Ausfälle und Verlangsamungen sind sofort sichtbar.

Grenze: Sie sagen nichts darüber, ob der Inhalt stimmt.

2 Form der Ausgabe

Anteil der Antworten, die das Schema verletzen: fehlende Pflichtfelder, unzulässige Werte.

Wirkung: der verlässlichste Frühindikator, weil er ohne bekanntes Sollergebnis auskommt.

Grenze: Ein Wert kann formal richtig und inhaltlich falsch sein.

3 Verteilung der Eingaben

Dokumentart, Sprache, Länge, Absender, Anteil gescannter Vorlagen.

Wirkung: Eine Verschiebung erklärt oft einen Qualitätsabfall, bevor jemand ihn meldet.

Grenze: Sie zeigt die Ursache, nicht die Wirkung.

4 Verteilung der Ausgaben

Antwortlänge, Anteil „nicht gefunden“, Häufigkeit der vergebenen Kategorien.

Wirkung: Wenn ein Modellwechsel wirkt, verschiebt sich hier meist zuerst etwas.

Grenze: Eine Verschiebung kann auch am veränderten Eingang liegen.

5 Rückmeldung aus dem Fachbereich

Wie oft ein Wert bei der Freigabe geändert wird, wie voll der Ausnahmekorb ist.

Wirkung: das einzige echte Qualitätssignal aus dem laufenden Betrieb.

Grenze: Es kommt verzögert und nur von den Stellen, an denen geprüft wird.

6 Verbrauch und Kosten

Aufrufe je Vorgang, Tokenmenge, Kosten je Vorgang — als Verteilung, nicht als Mittelwert.

Wirkung: Einzelne Ausreißer werden sichtbar, die im Durchschnitt untergehen.

Grenze: Steigende Kosten haben viele mögliche Ursachen.

Was die Schicht können muss

Sechs Punkte. Der teuerste Fehler ist, sie erst nach dem ersten Vorfall nachzurüsten.

■ Jeden Aufruf vollständig mitschreiben

Eingabe, Ausgabe, verwendetes Werkzeug, Dauer, Kosten. Auch die Läufe, die durchgegangen sind.

■ Versionen mitführen

Ohne Modell- und Promptversion am Datensatz lässt sich später kein Befund zuordnen.

■ Stichproben von Menschen prüfen lassen

Eine feste Quote zufällig gezogener Vorgänge. Nur so entstehen echte Qualitätszahlen im Betrieb.

■ Schwellen vorher festlegen

Ab welchem Wert jemand informiert wird, steht fest, bevor der erste Alarm ausgelöst wird.

■ Dem Fachbereich Zugang geben

Wer die Ergebnisse verantwortet, muss sie sehen können, ohne die IT zu fragen.

■ Aufbewahrung am Fachprozess ausrichten

Wo Ergebnisse nachweispflichtig sind, gilt die Frist des Fachprozesses, nicht die des Systems.

Vier Anlässe, bei denen jemand informiert wird

- Der Anteil der Schemaverletzungen steigt über die vereinbarte Schwelle.
- Der Anteil „nicht gefunden“ verändert sich deutlich gegenüber der Vorwoche.
- Die Kosten je Vorgang steigen, ohne dass das Volumen gestiegen ist.
- Die Modellversion hat gewechselt — das wird immer gemeldet, auch ohne Auffälligkeit.

Test und Monitoring beantworten verschiedene Fragen

Der Regressionstest fragt vor einer Änderung: Wird die neue Fassung schlechter als die alte?

Das Monitoring fragt im Betrieb: Verhält sich das System heute anders als letzte Woche?

Erst zusammen decken sie beide Richtungen ab: geplante Änderungen und schleichende Drift.

Wer nicht misst, erfährt von einer Verschlechterung durch eine Beschwerde.