DATALYXT

GRUNDLAGEN

Warum ein KI-System laufend geprüft wird

Software ändert sich, wenn jemand sie ändert. Ein KI-System ändert sich auch von allein.

Ein Regressionstest ist die einzige Stelle, an der auffällt, dass ein System schlechter geworden ist.

Was gemeint ist

Eine feste Menge echter Vorgänge, zu denen das richtige Ergebnis bekannt ist. Vor jeder Änderung läuft sie durch, und die Ergebnisse werden Feld für Feld mit den Sollwerten verglichen. Fällt die Trefferquote, wird die Änderung nicht eingespielt.

Was hier anders ist als bei klassischer Software

1. **Es gibt keine Zusicherung, nur eine Quote.**
   Ein Test in Programmcode ist bestanden oder nicht. Hier ist das Ergebnis eine Trefferquote über viele Fälle.

2. **Fehler sind still.**
   Ein Programm bricht ab. Ein Sprachmodell liefert einen plausibel aussehenden falschen Wert und läuft weiter.

3. **Eine Verbesserung an einer Stelle kann anderswo schaden.**
   Eine Prompt-Änderung wirkt auf alle Fälle zugleich, auch auf die, die vorher richtig waren.

4. **Die Grundlage verändert sich ohne Zutun.**
   Anbieter aktualisieren Modelle und stellen ältere Fassungen ab. Ein Wechsel kommt, ob geplant oder nicht.

5. **Auch die Eingaben driften.**
   Lieferanten ändern Layouts, Formulare bekommen neue Felder, Sprachen kommen hinzu.

Ohne Prüfmenge fällt eine Verschlechterung erst auf, wenn jemand aus dem Fachbereich sie meldet.

DATALYXT

ANFORDERUNGEN

Wie eine Prüfmenge aussehen muss

Sechs Eigenschaften. Fehlt eine davon, misst der Test etwas anderes als den Betrieb.

1 Echte Vorgänge, keine ausgedachten

Fälle aus dem Produktivbetrieb, anonymisiert übernommen.
Konstruierte Fälle prüfen das, was man erwartet hat — nicht das, was tatsächlich ankommt.
Umfang: lieber hundert echte als tausend erfundene.

2 Geprüfte Sollwerte je Feld

Zu jedem Fall steht fest, was richtig gewesen wäre, bestätigt von einer Fachkraft.
Sind sich zwei Fachkräfte uneinig, ist die Vorgabe unklar und nicht das Modell schlecht.
Ohne Sollwerte gibt es keinen Test, sondern nur einen Eindruck.

3 Feldweise gemessen

Für jedes Feld eine eigene Trefferquote statt einer Gesamtnote.
Ein Feld kann bei 99 Prozent liegen und das nächste bei 60. Der Durchschnitt verdeckt das.
Die schwachen Felder bestimmen, was der Fachbereich nachträglich prüfen muss.

4 Schwierige Fälle sind überrepräsentiert

Ausnahmen, schlechte Scans, fremde Sprachen, Werte an der Grenze.
Der Durchschnittsfall geht ohnehin durch und sagt wenig über die Belastbarkeit aus.
Jeder Fehler aus dem Betrieb wird als neuer Fall aufgenommen.

5 Stille Fehler werden getrennt gezählt

Unterschieden wird zwischen erkannt falsch und unerkannt falsch.
Erkannt falsch landet im Ausnahmekorb und kostet Bearbeitungszeit.
Unerkannt falsch läuft weiter. Nur diese Zahl ist gefährlich.

6 Sie wächst mit dem System

Neue Dokumentarten, neue Regeln und neue Lieferanten kommen als Fälle hinzu.
Eine eingefrorene Prüfmenge misst nach einem Jahr die Vergangenheit.
Wer sie pflegt, ist benannt — sonst pflegt sie niemand.

DATALYXT

BETRIEB

Wie der Lauf organisiert wird

Vier Schritte, vier Kennzahlen — und die vier Gründe, an denen es in der Praxis scheitert.

1 **Auslöser festlegen**

Jede Änderung an Prompt, Modell, Werkzeug oder Regelwerk startet den Lauf. Auch die des Anbieters.

2 **Im Schattenbetrieb laufen lassen**

Die neue Fassung verarbeitet dieselben Vorgänge mit, ohne das Produktivergebnis zu beeinflussen.

3 **Gegen vorher festgelegte Schwellen vergleichen**

Ab welchem Rückgang je Feld nicht eingespielt wird, steht vor dem Lauf fest.

4 **Freigeben und protokollieren**

Wer freigegeben hat, mit welchem Ergebnis, gegen welche Modell- und Promptversion.

Woran es in der Praxis scheitert

Die Prüfmenge entsteht erst nach dem ersten Vorfall

Für die Zeit davor gibt es dann keine Vergleichsgrundlage.

Es wird nur der Durchschnitt betrachtet

Ein einzelnes schwaches Feld verschwindet in einer guten Gesamtquote.

Der Lauf ist Handarbeit

Was manuell angestoßen wird, wird unter Termindruck ausgelassen.

Es fehlt eine Schwelle

Man sieht den Rückgang, aber niemand hat vorher festgelegt, ab wann er zählt.

Vier Kennzahlen, die je Lauf festgehalten werden

- Trefferquote je Feld, verglichen mit dem letzten freigegebenen Stand.
- Anteil stiller Fehler: plausibel aussehende, falsche Werte ohne Hinweis.
- Anteil Ausnahmekorb: Vorgänge, die zur Nachbearbeitung an Menschen gehen.
- Kosten und Laufzeit je Vorgang.

Die Prüfmenge entsteht vor dem ersten Produktivlauf, nicht nach dem ersten Vorfall.