DATALYXT

GRUNDLAGEN

Was passiert, wenn das Modell ausfällt

Ein Ersatzserver liefert dasselbe Ergebnis. Ein Ersatzmodell nicht.

Failover heißt bei Sprachmodellen: Ein anderes Modell übernimmt und löst dieselbe Aufgabe auf seine Weise.

Warum das kein gewöhnlicher Ausfallschutz ist

In der klassischen IT ist der Ersatz identisch: derselbe Programmstand auf einem zweiten Rechner, dasselbe Ergebnis. Bei Sprachmodellen gibt es diesen Ersatz nicht. Das zweite Modell formuliert anders, hält Formate anders ein und liegt an anderen Stellen daneben. Failover ist hier ein fachliches Thema.

Fünf Arten, wie ein Modell wegfällt

1. **Der Dienst ist nicht erreichbar.**
   Störung beim Anbieter, Netzwerkfehler, abgelaufene Zugangsdaten. Der Vorgang bleibt sofort stehen.

2. **Der Dienst drosselt.**
   Unter Last steigen Wartezeiten, Anfragen werden abgewiesen. Das System läuft, aber zu langsam.

3. **Die Modellfassung wird abgekündigt.**
   Anbieter stellen ältere Fassungen mit Frist ab. Der Wechsel kommt terminiert, aber er kommt.

4. **Das Modell verhält sich nach einem Update anders.**
   Formulierung, Ausführlichkeit und Formattreue verschieben sich, ohne dass etwas ausfällt.

5. **Der Zugang entfällt aus anderen Gründen.**
   Vertragslage, Preisänderung, regionale Einschränkungen oder Vorgaben zum Verarbeitungsort.

Nur der erste Fall ist ein Ausfall im gewohnten Sinn. Die anderen vier fallen ohne Vorsorge gar nicht auf.

DATALYXT

UNTERSCHIEDE

Andere Modelle, andere Probleme

Sechs Eigenschaften, in denen sich Modelle unterscheiden — und was beim Wechsel zuerst bricht.

1 Formattreue

Wie zuverlässig ein Modell eine vorgegebene Ausgabestruktur einhält.
Beim Wechsel bricht meist zuerst die Weiterverarbeitung, nicht die Antwort selbst.
Vorsorge: Die Schemaprüfung gilt für jedes Modell gleich und entscheidet über Annahme.

2 Kontextlänge

Wie viel Text in einen Aufruf passt.
Ein Ersatzmodell mit kleinerem Fenster zwingt dazu, Dokumente vorher aufzuteilen.
Vorsorge: Von Beginn an abschnittsweise arbeiten statt ganze Dokumente zu übergeben.

3 Ausführlichkeit und Ton

Ein Modell antwortet knapp, ein anderes erklärt zuerst und liefert die Angabe danach.
Nachgelagerte Textbausteine und Längenvorgaben passen dann nicht mehr.
Vorsorge: Aus Freitext nur das entnehmen, was das Schema vorsieht.

4 Sprachen und Fachvokabular

Die Qualität bei Deutsch, bei Fachbegriffen und bei fehlerhaft erkanntem Text ist verschieden.
Der Unterschied zeigt sich nicht im Mittel, sondern bei den schwierigen Vorgängen.
Vorsorge: Die Prüfmenge enthält genau diese Fälle und wird je Modell gemessen.

5 Werkzeugaufrufe

Wie zuverlässig ein Modell Funktionen aufruft und die Parameter richtig füllt.
Agentische Abläufe brechen hier als Erstes, weil jeder Schritt davon abhängt.
Vorsorge: Fehlerhafte Aufrufe abfangen und begrenzt wiederholen, statt sie zu übernehmen.

6 Betriebsrahmen

Verarbeitungsort, Vertragslage, Preis je Vorgang, Antwortzeit.
Das entscheidet oft, ob ein Ersatzmodell überhaupt eingesetzt werden darf.
Vorsorge: Die zulässigen Modelle stehen vorher fest und sind freigegeben.

DATALYXT

UMSETZUNG

Wie viel Vorsorge wann nötig ist

Vier Stufen. Die erste kostet fast nichts, wenn sie zu Beginn eingeplant wird.

0 Keine Vorsorge

Der Modellzugriff steht verstreut im Code. Fällt der Dienst aus, steht der Prozess.

1 Austauschbar

Der Zugriff liegt hinter einer eigenen Schnittstelle. Ein Wechsel ist eine Konfigurationsänderung.

2 Zweitmodell freigegeben

Ein Ersatzmodell ist gegen dieselbe Prüfmenge gemessen und fachlich abgenommen.

3 Automatische Umschaltung

Bei Fehlern oder Zeitüberschreitung übernimmt das Zweitmodell, sichtbar im Protokoll.

Welche Stufe für welchen Prozess

- Interner Assistent, ein Ausfall ist verschmerzbar: Stufe 1 genügt.
- Vorgänge mit Frist, etwa ein Rechnungslauf oder eine Meldung: Stufe 2.
- Durchlaufender Betrieb oder Kundenkontakt: Stufe 3.
- Wo kein Ersatz zulässig ist: sauber abbrechen und in den Ausnahmekorb geben.

Warum das von Anfang an gehört

Prompts hängen am Modell

Formulierungen, die bei einem Modell zuverlässig wirken, wirken beim nächsten anders.

Das Ausgabeformat hängt am Modell

Weiterverarbeitung und Werkzeugaufrufe brechen beim Wechsel als Erstes.

Schwellen und Prüfmenge hängen am Modell

Jede Kennzahl wurde gegen ein bestimmtes Modell gemessen und gilt für ein anderes nicht.

Nachträglich ist es ein Umbau

Wer die Austauschbarkeit später einzieht, fasst den Kern des Systems noch einmal an.

Ein zweites Modell gibt es immer. Die Frage ist, ob Ihr System es ohne Umbau verträgt.