Hilft ein Glossar gegen den KI-Drift? Acht Stufen, bis es wirkt

8. November 2026 · 12 Min.

> Lässt man die KI denselben Prozess mehrmals modellieren, heisst derselbe Schritt jedes Mal anders. Die naheliegende Antwort ist ein Glossar. Ich habe es gemessen, in vier Testreihen – und zuerst hat das Glossar rein gar nichts gebracht.

In der letzten Folge habe ich gezeigt: Lässt man die KI denselben Prozess mehrmals modellieren, kommt jedes Mal etwas anderes heraus. Eine einzige Kundenbestellung hatte 32 verschiedene Namen. Die naheliegende Antwort darauf ist ein Glossar. Man sagt der KI, welche Wörter gelten, und das Problem ist gelöst.

Genau das habe ich gemacht. Und ich war erstaunt, als das Glossar zuerst rein gar nichts brachte. Dieser Artikel fasst zusammen, warum, und was es braucht, damit ein Glossar wirklich wirkt. Das Video zur Folge zeigt dieselbe Geschichte mit den Bildschirmen aus dem Labor (deutsche Fassung, englische Fassung). Die Unterlagen und das Glossar aus dem Versuch gibt es am Ende zum Herunterladen.

Was Drift ist, und warum er nicht nur Zufall ist

Drift heisst: dieselben Unterlagen, dieselbe Bitte an das Sprachmodell, mehrere Läufe. Die Struktur bleibt oft ähnlich, aber die Wörter wandern. Ein Schritt heisst einmal «Check order», dann «Validate order», dann «Verify order».

Ein Teil davon ist gewollt. Sprachmodelle würfeln ein wenig, wenn sie das nächste Wort wählen; das steuert die Temperatur. Man kann sie auf null setzen. Ich habe das getan, und trotzdem kamen in meinen Vorversuchen jedes Mal andere Unterprozesse heraus. Das ist kein Einzelfall: Auch bei Temperatur null liefern Sprachmodelle auf dieselbe Frage nachweislich unterschiedliche Antworten (Ouyang et al., 2025). Wer Stabilität will, kann sie also nicht einfach einschalten.

Der grössere Teil des Drifts hat ohnehin eine andere Ursache. Er steckt in den Unterlagen.

Der Fall: 4Health

Der Versuch läuft am Projekt «Sales & Order Management» der fiktiven Firma 4Health, einem Hersteller von Sportriegeln mit Online-Shop, Riegel-Konfigurator und Handelspartnern, mit Standorten in fünf Ländern. Der Prozess heisst «B2C order-to-cash», von der Bestellung im Shop bis zur Zahlung.

Dazu gehören 28 Dokumente, wie sie in jeder Firma herumliegen: eine Prozessbeschreibung, die AGB, eine Schnittstellenbeschreibung zum ERP, Interviewnotizen, E-Mails aus München und Italien,

eine Schulung von 2023, Excel-Listen. Und wie in jeder Firma widersprechen sie sich, vor allem in ihren Wörtern. Die Prozessbeschreibung sagt «Order», die deutsche E-Mail «Bestellung», die französische Anleitung «commande client», das ERP «Sales Order». Für die Gutschrift gibt es sechs Wörter, für die Mahnung sieben.

Die KI macht also nichts falsch. Sie wählt bei jedem Lauf ein anderes Wort, weil niemand entschieden hat, welches gilt. Genau dafür ist ein Glossar da: eine Liste von Begriffen, nicht von Wörtern. Die Terminologielehre trennt diese beiden Ebenen seit Langem: Ein Begriff ist die Sache, die Benennungen sind die Wörter dafür, und ein Begriff kann mehrere haben (ISO 704:2022).

Ohne Glossar: 190 Namen

Zuerst die Ausgangslage. Die KI baut alles selbst: zuerst die Analyse mit den Aktivitäten, dann das Modell mit allen Unterprozessen, rund 50 Aktivitäten pro Lauf. Fünf Läufe, ohne Glossar, nur mit den Unterlagen.

Über fünf Läufe kamen **190 verschiedene Aktivitätsnamen** zusammen. Allein rund um die Rechnung gab es 17 verschiedene Aktivitäten. «Create invoice», «Generate invoice», «Invoice order» und «Create invoice document» meinen dieselbe Arbeit.

Und beim ersten Wort: Die KI hat in diesen fünf Läufen **65 verschiedene Verben** benutzt. Capture oder receive, check oder validate, create oder issue, für dieselben Handgriffe.

Acht Stufen, bis das Glossar wirkt

Es hat acht Stufen gebraucht, bis das Glossar etwas bewirkt hat. Die ersten fünf haben es aufgebaut. Die Zahlen haben sich dabei kaum bewegt.

Stufe	Was fehlte	Was es brauchte
1 · Wörterliste	Eine Liste sagt, welche Wörter es gibt, aber nicht, was sie bedeuten, und nicht, welche falsch sind.	Ein echtes Glossar mit Bedeutung, Varianten und Sprachen
2 · Altes Glossar	Das vorhandene Glossar stammte aus einem Projekt zur Einführung von BPM: viele gute Wörter, nur nicht die der Bestellabwicklung.	Die Begriffe des Fachgebiets
3 · KI-Vorschläge	Von Hand schreibt niemand ein Glossar aus über hundert Dokumenten in vier Sprachen. Die KI fand 2072 Vorschläge, aber einen pro Wort, nicht pro Begriff.	Eine Zusammenführung
4 · Kuration	Aus fünf Zeilen für die Gutschrift muss ein Begriff werden, mit einem Namen pro Sprache und den anderen Wörtern als Synonyme.	Zuordnung jedes Vorschlags zu einem Begriff, mit Fokus auf das Projekt
5 · Bestätigen	Alle Begriffe waren Entwürfe, und Entwürfe gehen nicht an die KI. Eine Testreihe war darum ungültig.	Fachleute, die bestätigen und ablehnen
6 · Ankommen	Von 473 Begriffen kamen genau drei bei der KI an.	Synonyme, die einen Begriff in den Prompt holen
7 · Verben	Das Glossar regelt Nomen. Der Drift steckt in den Verben.	Eine Verbregel
8 · Kernobjekte	Die drei Dinge, um die sich der Prozess dreht, fehlten im Glossar.	Von Hand ergänzt

Drei Stufen verdienen einen genaueren Blick.

Stufe 6: Das Glossar kommt nicht an

Reihe B lief mit 484 bestätigten Begriffen. Ergebnis: 181 statt 190 verschiedene Namen. Praktisch nichts.

Also habe ich nachgesehen, was wirklich bei der KI ankommt. Mein Labor-Werkzeug gibt ihr nicht das ganze Glossar mit, sonst wird die Anfrage zu lang, sondern nur die Begriffe, die der Prozess erwähnt. Und erwähnt hiess: Das Wort steht wörtlich im Text des Prozesses. Steht dort «credit memo», heisst der Begriff aber «credit note», kommt er nicht mit. Synonyme zählten nicht. Von 473 Begriffen kamen genau drei an: «invoice», «online shop» und «INVOIC», der Name einer EDI-Nachricht.

Seit der Korrektur zählen die Synonyme mit, und es kommen 31 Begriffe an, die zu diesem Prozess passen. Und noch etwas: Das muss schon in der Analyse passieren. Dort wählt die KI zum ersten Mal, wie die Schritte heissen. Das Design übernimmt die Wörter nur noch.

Die Lehre ist unbequem, weil sie nicht in der Benutzeroberfläche sichtbar ist. Ein Glossar kann vollständig, kuratiert und bestätigt sein und trotzdem nichts bewirken, wenn es den Weg in die Anfrage nicht findet.

Stufe 7: Die Verben

Auch ein Glossar, das ankommt, hat eine Lücke. Es regelt Nomen: was eine Bestellung ist, was eine Gutschrift. Ein Aktivitätsname besteht aber aus Verb und Objekt, so verlangen es die Modellierungsrichtlinien seit Langem (Silver, 2011; Mendling, Reijers & Recker, 2010). Und beim Verb driftet die KI am stärksten.

Also bekommt die KI zusätzlich eine kurze Verbliste: ein verbindliches Verb je Art von Arbeit, «Check» statt «Validate» oder «Verify», «Create» statt «Generate» oder «Issue». Und das Werkzeug korrigiert das Verb einer Aktivität, wenn es ausserhalb der Liste liegt. Die Idee ist nicht neu; die Forschung zur automatischen Korrektur von Aktivitätsnamen arbeitet genau mit dieser Zerlegung in Verb und Objekt (Leopold, Smirnov & Mendling, 2012).

Reihe B2: statt 65 nur noch **13 verschiedene Verben**, und die Namen gehen von 190 auf **135** zurück. Der erste wirklich grosse Schritt.

Mit einer Nebenwirkung: Plötzlich hiess vieles einfach «Create». «Create shipment» statt «Ship order». Aus einer fachlichen Tätigkeit wird das Anlegen eines Belegs. Darum erlaubt die Liste jetzt ausdrücklich fachliche Verben, die in diesem Prozess Sinn ergeben: ship, produce, pick, pack.

Stufe 8: Die Kernobjekte

Mit den Verben im Griff wanderte der Drift weiter, zu den Objekten. Heisst es «ERP order» oder «sales order»? «Production» oder «production order»? «Shipment» oder «order»?

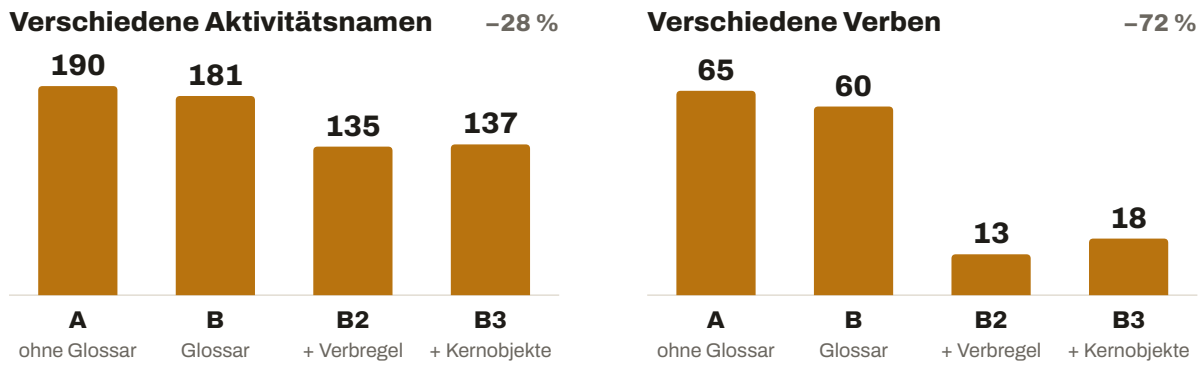
Genau hier hätte das Glossar helfen sollen. Ich habe nachgesehen: Diese Begriffe fehlten. Das kuratierte Glossar hatte 484 Begriffe, aber nicht die drei Dinge, um die sich dieser Prozess dreht. Also habe ich sie ergänzt: «sales order», «production order» und «parcel», und «ERP order» als abgelehnten Begriff festgelegt.

Reihe B3: «ERP order» kommt in keinem Lauf mehr vor, «shipment» nur noch zweimal. Die Zahl der Namen bleibt bei 137, aber jetzt sind es die richtigen Wörter.

Die Lehre: Die Kernobjekte eines Prozesses müssen im Glossar stehen. Eine KI, die tausende Begriffe findet, findet nicht unbedingt die drei wichtigsten.

Was sich ändert

Vier Testreihen zu je fünf Läufen, derselbe Prozess, dasselbe Modell, dieselben Unterlagen:



Verschiedene Aktivitätsnamen und verschiedene Verben über je fünf Läufe. Den grössten Sprung macht Reihe B2, die Verbregel.

Die Zahlen im Einzelnen:

Reihe	Glossar	verschiedene Namen	verschiedene Verben	Ereignisse in allen 5 Läufen gleich	Flüsse in allen 5 Läufen gleich
A	ohne	190	65	6	1
B	484 bestätigte Begriffe	181	60	6	3
B2	dazu Synonyme und Verbregel	135	13	7	4
B3	dazu die Kernobjekte	137	18	10	8

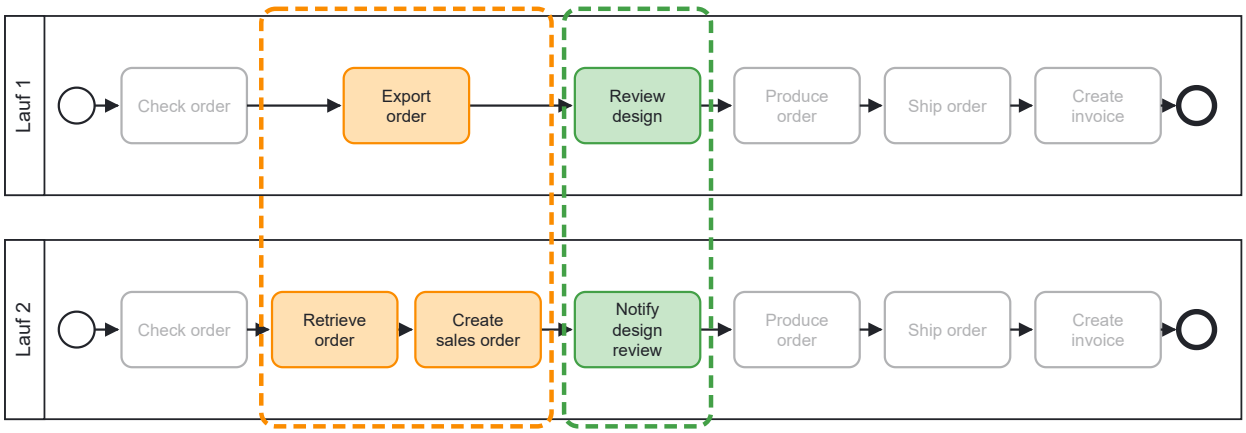
Die Namen gehen um gut ein Viertel zurück, die Verben von 65 auf 18. Der grösste Effekt kam also nicht vom Glossar, sondern von der Verbregel daneben. Und was in allen fünf Läufen gleich bleibt, nimmt zu: zehn statt sechs Ereignisse, acht statt einem Fluss zwischen den Schritten.

Dazu ist jede Korrektur nachvollziehbar: welches Wort ersetzt wurde, wodurch, und aufgrund welches Glossareintrags.

Ist es damit stabil? Nein, nicht ganz. Aber deutlich ruhiger, und die Wörter, die bleiben, sind die der Firma.

Was das Glossar nicht löst

Erstens den Zuschnitt der Schritte. Auf der obersten Ebene gab es ohne Glossar sogar am wenigsten Bewegung: neun verschiedene Namen, fünf davon in allen Läufen gleich. Mit Glossar waren es sechzehn. Dort ändert sich nicht mehr das Wort, sondern der Schnitt.



Zwei der fünf Läufe aus Reihe B3, oberste Ebene. Blass: in beiden gleich. Orange: Die Übergabe an das ERP ist einmal ein Schritt, einmal zwei. Grün: Einmal prüft der Prozess das Design selbst, einmal stösst er die Prüfung nur an.

Ist die Übergabe an das ERP ein eigener Schritt oder zwei? Prüft der Prozess das Design selbst, oder stösst er die Prüfung nur an? Das ist keine Frage der Wörter. Das muss eine bestätigte Analyse festhalten, damit die KI es nicht bei jedem Lauf neu erfindet. Schon die klassischen Grundsätze ordnungsmässiger Modellierung trennen die Einheitlichkeit der Begriffe von der Frage, was ein Modell abbildet und wie fein (Becker, Rosemann & von Uthmann, 2000).

Zweitens die Arbeit der Fachleute. Die KI findet die Begriffe und schlägt Gruppen vor, aber sie macht Fehler, und nicht jedes Mal dieselben. Und, wie wir gesehen haben, sie übersieht ausgerechnet die wichtigsten. Bestätigen und ergänzen muss ein Mensch.

Selbst ausprobieren

Auf bpmn.ch gibt es das Material aus dem Versuch zum Herunterladen (bpmn.ch/v/2): die 28 Dokumente von 4Health, das Glossar mit 487 bestätigten Begriffen samt Synonymen, abgelehnten Wörtern und Fundstellen, die Verbliste, eine Anleitung und ein Messblatt.

Damit lässt sich der Kern des Versuchs mit jedem Werkzeug wiederholen: fünfmal ohne Glossar, fünfmal mit Glossar und Verbliste, und dann zählen, wie viele verschiedene Namen und Verben herauskommen. Die Zahlen aus einem Chat sind nicht direkt mit denen aus dem Video vergleichbar, weil dort ein ganzes Modell mit Unterprozessen entstand. Der Unterschied zwischen den beiden Reihen ist es schon.

Das Glossar liegt bewusst so bei, wie es im Versuch verwendet wurde, nicht nachpoliert. Wer es durchsieht, findet auch die Fehler der KI, und damit den besten Beleg dafür, warum am Ende jemand vom Fach bestätigen muss.

MESTRO, mein Labor-Werkzeug, setzt die Stufen 3 bis 8 um: KI-Vorschläge mit Fundstelle, den Kurationsvorschlag, die Bestätigung, die Synonyme im Prompt schon bei der Analyse, die Verbregel und die nachvollziehbare Korrektur. Es dient hier als Beispiel dafür, wo die Arbeit liegt, nicht als Produktvergleich.

Fazit

Hilft ein Glossar gegen Drift? Allein: fast nichts. Ein volles, bestätigtes Glossar hat die Zahl der Namen in meinem Versuch kaum bewegt, weil es nicht bei der KI ankam.

Es wirkt, wenn mehreres zusammenkommt. Die Begriffe kommen samt Synonymen bei der KI an. Es gibt eine Regel für die Verben, nicht nur für die Nomen. Und die Kernobjekte des Prozesses stehen drin. Dann gehen die Namen um gut ein Viertel zurück und die Verben von 65 auf 18.

Ein Glossar ist also nötig, aber nicht hinreichend. Was es nicht regelt, ist der Zuschnitt der Schritte. Dafür braucht es eine Analyse, die jemand bestätigt hat.

Die gute Nachricht: Den Grossteil der Arbeit macht die KI. Sie liest die Unterlagen, findet die Begriffe und schlägt Gruppen vor. Die Entscheidung aber liegt beim Menschen.

Und eine Frage an euch: Wie viele Wörter hat eure Firma für «Bestellung»?

Hinweis zum Versuch: Prozess «B2C order-to-cash» der fiktiven Firma 4Health, Modell gpt-oss, Ausgabe Englisch, je fünf Läufe pro Reihe, November 2026. Pro Lauf baut die KI die Analyse neu, danach das Modell mit allen Unterprozessen; ausgewertet automatisch über alle Aktivitäten, Ereignisse und Flüsse. Die Reihen B2 und B3 liefen mit einem neueren Stand des Werkzeugs (Synonyme im Prompt, Verbregel). Alle Unterlagen sind erfunden. MESTRO ist mein Labor-Werkzeug und entsteht bei ELCA, wo ich arbeite; es dient hier als Referenz-Implementierung, nicht als Produktvergleich.

Quellen & Weiterlesen

- Video zur Folge 2 «Hilft ein Glossar gegen den KI-Drift?» auf dem Kanal «Stephan Fischli · Agentic Orchestration»: deutsche Fassung, englische Fassung
- Unterlagen und Glossar aus dem Versuch auf bpmn.ch: bpmn.ch/v/2
- Folge 1 «Text rein, Modell raus?», dazu der Artikel «Text rein, Modell raus? Fünf Stufen, an denen die KI scheitert».
- J. Becker, M. Rosemann, C. von Uthmann: *Guidelines of Business Process Modeling*. In: W. van der Aalst, J. Desel, A. Oberweis (Hrsg.): Business Process Management, LNCS 1806, Springer, 2000, S. 30–49.
- ISO 704:2022: *Terminology work — Principles and methods*. International Organization for Standardization, 2022.
- H. Leopold, S. Smirnov, J. Mendling: *On the refactoring of activity labels in business process models*. Information Systems 37(5), 2012, S. 443–459.
- J. Mendling, H. A. Reijers, J. Recker: *Activity labeling in process modeling: Empirical insights and recommendations*. Information Systems 35(4), 2010, S. 467–482.

- S. Ouyang, J. M. Zhang, M. Harman, M. Wang: *An Empirical Study of the Non-determinism of ChatGPT in Code Generation*. ACM Transactions on Software Engineering and Methodology 34(2), 2025.
- B. Silver: *BPMN Method and Style*, 2. Auflage, Cody-Cassidy Press, 2011. https://methodandstyle.com/