Text rein, Modell raus? Fünf Stufen, an denen die KI scheitert

17. Oktober 2026 · 12 Min.

> Man beschreibt den Prozess im Chat, und die KI zeichnet das BPMN-Modell. Ich habe diese Erwartung gemessen: fünf Sprachmodelle, drei Arten von Unterlagen, über 170 generierte Modelle. Das Ergebnis ist eine Messlatte mit fünf Stufen und eine These, die sich als falsch erwiesen hat.

Text rein, Modell raus. So stellen sich die meisten die Prozessmodellierung mit KI vor, und ich am Anfang auch. Man beschreibt den Prozess in ein paar Sätzen im Chat, und das Sprachmodell liefert das fertige BPMN-Modell. Wenn das stimmt, wozu braucht es dann noch Methode, Stilregeln und Standards?

Ich habe diese Frage nicht diskutiert, sondern gemessen: mit fünf Sprachmodellen, drei Arten von Unterlagen und über 170 generierten Modellen, jedes einzelne automatisch geprüft. Dieser Artikel fasst den Versuch und seine Ergebnisse zusammen. Das Video dazu zeigt dieselbe Geschichte mit den echten Modellen aus dem Labor (deutsche Fassung, englische Fassung). Und wer es selbst nachmessen will, findet am Ende das Material dazu.

Die Gegenposition, in ihrer stärksten Form

Bevor ich zum Versuch komme, die Gegenposition. Sie verdient es, ernst genommen zu werden.

Erstens: Die Sprachmodelle kennen BPMN. Sie haben die Spezifikation gelesen und Tausende von Modellen gesehen. **Zweitens:** Sie werden jedes Jahr besser; was heute hakt, ist morgen gelöst. **Drittens:** Wer mehr Unterlagen mitgibt, bekommt ein besseres Modell; mehr Kontext heisst mehr Wissen. **Viertens:** Methode und Stilregeln sind Ballast aus einer Zeit, in der Menschen noch selbst gezeichnet haben.

Manches davon stimmt. Die Modelle kennen die Notation tatsächlich, und sie werden besser. Die Frage ist nur, ob das genügt.

Warum der Chat allein nicht reicht

Wer es selbst ausprobiert, erlebt meistens dasselbe. Ein starkes Modell liefert ein Diagramm, das sich öffnen lässt und plausibel aussieht. Der Spur nach richtig. Brauchbar ist es oft nicht.

Der Grund liegt nicht in mangelnder Intelligenz, sondern in fehlendem Rahmen. Der Chat kennt das Unternehmen nicht, seine Begriffe nicht, seine Regeln nicht. Also tut er, was Sprachmodelle tun: Er sucht nach dem wahrscheinlichsten Muster, er interpretiert, und wo etwas fehlt, ergänzt er. Plausibel,

aber erfunden. Die Forschung nennt das Halluzination, und sie ist kein Fehler einzelner Modelle, sondern eine Eigenschaft der Technik (Ji et al., 2023).

Gefährlich wird das, weil das Ergebnis gut aussieht. Die Automationsforschung beschreibt seit Jahrzehnten, dass Menschen einem maschinellen Ergebnis umso eher folgen, je zuverlässiger es wirkt (Parasuraman & Riley, 1997). Ein sauber gezeichnetes, aber unvollständiges Modell ist deshalb riskanter als ein offensichtlich kaputtes.

Und wer dann die vorhandenen Unterlagen hochlädt, das Word-Dokument, die Schulungsfolien, eine Notiz, macht es nicht unbedingt besser. Unterlagen widersprechen sich, wie in jedem echten Projekt. Die KI entscheidet dann still, welcher sie glaubt.

So reden alle aneinander vorbei. Die einen sagen: funktioniert doch. Die anderen: alles halluziniert. Beide haben ein Beispiel. Was fehlt, ist eine Messlatte.

Eine Messlatte: fünf Stufen

Ich schlage fünf Stufen vor, die aufeinander aufbauen. Jede setzt die vorherige voraus.

Stufe	Name	Die Frage
1	**Es lädt**	Ist das Ergebnis gültiges BPMN, mit Diagramm, in jedem Werkzeug zu öffnen?
2	**Vollständig**	Kommen alle beschriebenen Fälle vor, nicht nur der Normalfall, auch die Ausnahmen?
3	**Stabil**	Wer zweimal fragt, bekommt der zweimal denselben Ablauf?
4	**Treu**	Ist nichts erfunden, und bleiben die Begriffe die Begriffe der Fachleute?
5	**Ehrlich**	Wo sich die Quellen widersprechen: Fragt die KI, statt still zu entscheiden?

Erst ab Stufe 4 kann man einem Modell vertrauen. Und erst mit Stufe 5 kann man es verantworten.

Der Versuch

Der Prozess ist bewusst alltäglich: eine Retoure im Onlinehandel. Der Kunde meldet eine Rückgabe an, sie wird geprüft, die Ware kommt zurück, wird kontrolliert, und das Geld wird erstattet. Oder eben nicht: Die Rückgabe wird abgelehnt, oder die Ware ist nicht in Ordnung.

Als Eingabe dienten drei Varianten: ein sauberer Text; eine Notiz aus dem Gedächtnis, so wie sie in der Praxis entsteht; und drei Unterlagen, die sich bewusst ein wenig widersprechen (ein Word-Dokument, Schulungsfolien und diese Notiz). Jede Variante ging zehnmal an jedes Sprachmodell, immer mit derselben einfachen Bitte: «Erstelle aus dieser Beschreibung ein BPMN-2.0-Modell als XML.» Eine Nachricht, keine Systemanweisung, kein Nachbessern. Jedes Ergebnis wurde gespeichert und automatisch gegen die Stufen geprüft.

Stufe 1: Lädt es?

Mit der Notiz als Eingabe lieferten drei von fünf Modellen in keinem einzigen Lauf ein Modell, das sich öffnen lässt.

Modell	gültig, mit Diagramm (von 10)
gpt-oss (120b)	0
DeepSeek V4 Flash	0
GPT-4.1	0
GPT-5 mini	3
GLM-5	10

Die Fehler sind lehrreich. Mal ist das XML schlicht kaputt, etwa weil ein Kommentar einen doppelten Bindestrich enthält, was XML verbietet. Häufiger fehlt der Diagramm-Teil ganz: Das Modell beschreibt einen Prozess, aber es zeichnet keinen. Ein Viewer zeigt dann nur «no diagram to display».

Die erste Stufe ist also nicht selbstverständlich. Und sie ist reine Technik; über den Prozess selbst sagt sie noch gar nichts.

Stufe 2 und 3: Vollständig und stabil?

Die Notiz sagt klar, dass das Lager prüft, ob die Ware stimmt und in Ordnung ist. Ein vollständiges Modell braucht also einen Pfad für den Fall, dass sie es nicht ist.

GLM-5, das Modell, das in Stufe 1 immer gültig gezeichnet hat, modellierte diesen Pfad in einem von zehn Läufen. Die anderen neun zeigen den Normalfall, sauber gezeichnet. Genau das ist das Gefährliche: ein Modell, das gut aussieht, aber nur den Normalfall kennt.

Und zur Stabilität: Zehnmal dieselbe Frage ergab bei GLM-5 sechs verschiedene Abläufe, bei GPT-5 mini sieben. Nicht andere Namen, sondern andere Schritte, andere Reihenfolgen, andere Ausgänge.

Hier kommen die Standardwerkzeuge nicht weiter. Ein besseres Modell hilft bei Stufe 1. Bei 2 und 3 hilft es nicht.

Was die Methode ändert

Im nächsten Schritt lief derselbe Versuch durch eine Methode, umgesetzt in meinem Labor-Werkzeug MESTRO. Bewusst mit dem schwächsten Modell aus dem Versuch, gpt-oss, demjenigen, das allein kein einziges gültiges Modell geschafft hatte.

Methode heisst hier vier Dinge. Die KI füllt ein festes **Schema**, statt frei XML zu schreiben. Sie folgt **Stilregeln** aus *BPMN Method and Style* (Silver, 2011): Aktivitäten als Verb und Objekt, benannte Endzustände, Entscheidungen als Fragen. Das Layout macht eine **Konvention**, nicht das Sprachmodell. Und jedes Ergebnis wird **geprüft**, bevor ein Mensch es sieht; was die Regeln verletzt,

geht zurück an die KI. Das sind keine neuen Ideen: Die empirische Modellierungsforschung hat solche Leitlinien längst als Treiber von Verständlichkeit und Fehlerarmut belegt (Mendling, Reijers & van der Aalst, 2010). Auch die Prozessforschung, die Sprachmodelle Prozessmodelle erzeugen lässt, setzt auf eine feste Zwischenstruktur und eine Prüfschleife statt auf freie Ausgabe (Kourani et al., 2024).

Das Ergebnis: zehn von zehn Modellen gültig, mit Diagramm. Der Pfad «Ware nicht in Ordnung» war in neun bis zehn von zehn Läufen da.

Aber ehrlich: Stabil ist es nicht. Auch mit Methode entstanden fünf bis sieben verschiedene Abläufe. Die Methode macht sicher, was beschrieben ist. Was offen ist, bleibt offen.

Stufe 4: Mehr Unterlagen, mehr Wahrheiten

Für Stufe 4 hatte ich eine These: Je mehr Unterlagen, desto mehr Fälle findet die KI. Die Unterlagen enthalten zum Beispiel eine Frist, die in der Notiz fehlt: Die Ware muss innerhalb von sieben Tagen zurück sein.

Das Gegenteil ist passiert. Mit den drei Unterlagen gab es bei fast allen Modellen zehn verschiedene Abläufe in zehn Läufen, auch mit Methode. Die Sieben-Tage-Frist tauchte in höchstens drei von zehn Läufen auf.

Und die Widersprüche? Die Notiz sagt das eine, die Folien und das Word-Dokument das andere:

Notiz	**Folien / Word**
nur Rückerstattung	oder Ersatzlieferung
kein formaler Abschluss	Fall wird geschlossen
kein Supervisor	Supervisor beiziehen

Jedes Modell entschied still für eine Seite. Die Ersatzlieferung landete bei allen Modellen in acht bis zehn von zehn Läufen im Modell. Und aus dem «Refund Request», den die Fachleute in der Notiz ausdrücklich so nennen, wurde mit Methode in zehn von zehn Läufen eine «Rücksendeanfrage».

Die These war falsch. Mehr Unterlagen bringen nicht mehr Fälle. Sie bringen mehr Wahrheiten. Und die KI wählt eine davon, ohne es zu sagen.

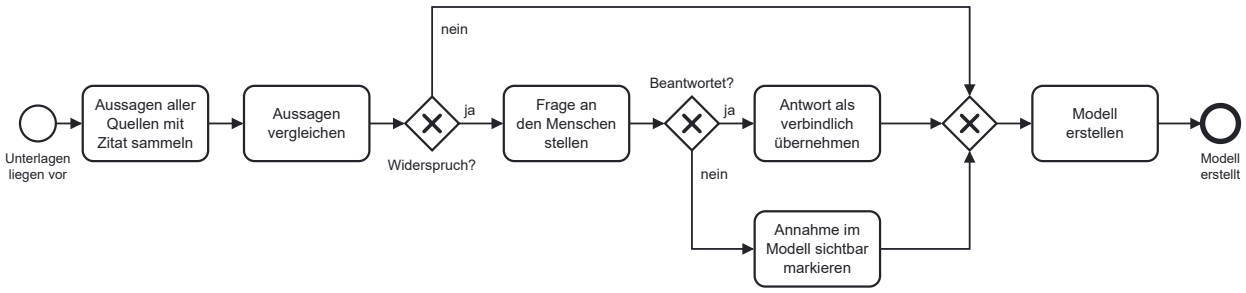
Stufe 5: Die KI sieht den Widerspruch und löst ihn still

Der aufschlussreichste Befund steckt in den Rohantworten der reinen Chats. DeepSeek schrieb in die Dokumentation seines Modells: «Quellen weichen ab bzgl. Eskalationen, Ersatzlieferungen und Begriffsverwendung.» gpt-oss schrieb: «Ersetzen vs. Refund – Beide Optionen werden … dargestellt; das Unternehmen kann … entscheiden.»

Die Sprachmodelle sehen die Widersprüche also. Sie lösen sie nur selbst, mal so, mal so. Genau daher kommt der Drift aus Stufe 4.

In meinem Labor war es nicht anders. Die Qualifizierung des Prozesses nannte die Widersprüche sogar wörtlich, etwa «Unstimmigkeit: Ersatzlieferung vs. reine Rückerstattung». Aber dieses Wissen floss nirgends hin; das generierte Modell enthielt die Ersatzlieferung trotzdem.

Also haben wir nachgebessert. Nicht am Sprachmodell: Das ist dasselbe geblieben, trainiert wurde nichts. Geändert haben wir den Ablauf.



Bevor modelliert wird, werden die Aussagen aller Quellen mit ihrem Zitat gesammelt und verglichen. Jeder Widerspruch wird zur Frage an den Menschen. Seine Antwort ist dann verbindlich. Und was niemand beantwortet, steht als sichtbare Annahme im Modell, statt still entschieden zu werden. Wie weit dieser Quellen-Abgleich Stufe 5 tatsächlich erreicht, messe ich gerade nach; das Ergebnis folgt.

Was bleibt, Stufe für Stufe

Stufe	Befund
1 · Es lädt	Schafft ein starkes Sprachmodell allein. Mit Methode schafft es auch ein schwaches.
2 · Vollständig	Schafft erst die Methode. Der Chat zeichnet den Normalfall.
3 · Stabil	Schafft niemand ganz, auch die Methode nicht. Wo die Beschreibung offen ist, entscheidet die KI jedes Mal neu.
4 · Treu	Mit mehreren Unterlagen scheitern alle, solange die KI Widersprüche selbst löst.
5 · Ehrlich	Erst, wenn aus jedem Widerspruch eine Frage wird.

Die Stufen 3, 4 und 5 hängen zusammen. Stabil und treu kann ein Modell erst werden, wenn jemand die offenen Stellen entscheidet. Und dieser Jemand ist nicht die KI.

Selbst nachmessen

Die Stufen sind einfach genug, um sie mit jedem Werkzeug nachzumessen, ohne ein Programm. Auf bpmn.ch gibt es dazu ein Paket: die Notiz aus dem Versuch, die Aufgabe, eine Prüfliste für die Stufen 1 bis 3, ein Messblatt und die Vergleichswerte aus dem Video (bpmn.ch/v/1). Zehnmal fragen, jedes Ergebnis in einem BPMN-Viewer öffnen, abhaken.

Warum die Prüfliste bei Stufe 3 aufhört. Weil bis dorthin jeder mitmessen kann, mit jedem Werkzeug und jedem Modell. Lädt es? Ist der Ausnahmepfad da? Kommt zweimal dasselbe heraus? Das sieht man im Viewer und hakt es auf dem Messblatt ab. Mehr braucht es nicht — kein Zugang, keine Programmierung, keine Absprache mit mir. Genau deshalb sind diese drei Stufen die Messlatte: Sie sind für alle dieselbe.

Ab Stufe 4 ändert sich die Art der Frage. Treue lässt sich erst prüfen, wenn mehrere Quellen vorliegen, die sich widersprechen — man muss den Fall also erst aufbauen. Und Stufe 5 lässt sich am Modell überhaupt nicht ablesen. Sie fragt nicht, wie das Ergebnis aussieht, sondern ob das Werkzeug **zurückgefragt** hat. Ein Werkzeug, das still entschieden hat, liefert ein tadelloses Diagramm; der Fehler ist im Ergebnis unsichtbar. Sehen kann man ihn nur im Ablauf davor.

Damit hört es auf, eine Eigenschaft des Sprachmodells zu sein, und wird eine Eigenschaft des Rahmens drumherum — des Quellen-Abgleichs mit der Frage an den Menschen, den ich oben beschrieben habe. Im Versuch habe ich diesen Schritt in MESTRO nachgerüstet; ob er Stufe 5 wirklich erreicht, messe ich noch.

MESTRO setzt an allen fünf Stufen an, und zwar jeweils dort, wo die Stufe entschieden wird: Schema und Prüfschleife bei Stufe 1, die Methode bei Stufe 2, Stilregeln und Layout-Konvention bei Stufe 3, der Quellen-Abgleich bei Stufe 4 und die Frage an den Menschen bei Stufe 5. Ansetzen heisst nicht erreichen — die Stufen 3 und 5 sind in diesem Versuch noch offen. Aber es zeigt, wo die Arbeit liegt: nicht am Sprachmodell, sondern am Rahmen.

Wer trotzdem selbst daran will: Die drei sich widersprechenden Unterlagen liegen im Paket mit dabei. Für Stufe 4 reicht das. Für Stufe 5 ist die Frage weniger, was Ihr Werkzeug zeichnet, als ob es Sie etwas fragt.

Mich interessiert, welche Stufe euer Werkzeug erreicht, und wo es bei euch bricht.

Fazit

Die Gegenposition hat in einem Punkt recht: Die Sprachmodelle können BPMN, und die besten zeichnen heute zuverlässig gültige Modelle. Aber Stufe 1 ist Technik, nicht Prozessverständnis. Vollständigkeit kam im Versuch erst mit Methode. Und mehr Unterlagen machten die Modelle nicht besser, sondern beliebiger, weil die KI Widersprüche, die sie selbst erkennt, still für sich entscheidet. Verlässlichkeit kommt nicht aus dem Sprachmodell. Sie kommt aus dem Rahmen, in dem es arbeitet, und aus der Frage an den Menschen.

Eine KI, die nicht fragt, entscheidet für euch.

Hinweis zum Versuch: Abruf der Modelle über die Programmierschnittstelle mit Standard-Temperatur, je zehn Läufe pro Eingabe, Stand Oktober 2026. Ausgewertet automatisch; alle Unterlagen sind erfunden. MESTRO ist mein Labor-Werkzeug und entsteht bei ELCA, wo ich arbeite; es dient hier als Referenz-Implementierung der Methode, nicht als Produktvergleich.

Quellen & Weiterlesen

- Video zur Folge 1 «Text rein, Modell raus?» auf dem Kanal «Stephan Fischli · Agentic Orchestration»: deutsche Fassung, englische Fassung
- Nachmess-Paket zum Video auf bpmn.ch: bpmn.ch/v/1
- Z. Ji, N. Lee, R. Frieske et al.: *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys 55(12), 2023.
- H. Kourani, A. Berti, D. Schuster, W. M. P. van der Aalst: *Process Modeling with Large Language Models*. In: Enterprise, Business-Process and Information Systems Modeling (BPMDS/EMMSAD 2024), LNBIP, Springer, 2024.
- J. Mendling, H. A. Reijers, W. M. P. van der Aalst: *Seven process modeling guidelines (7PMG)*. Information and Software Technology 52(2), 2010, S. 127–136.
- R. Parasuraman, V. Riley: *Humans and Automation: Use, Misuse, Disuse, Abuse*. Human Factors 39(2), 1997, S. 230–253.
- B. Silver: *BPMN Method and Style*, 2. Auflage, Cody-Cassidy Press, 2011. https://methodandstyle.com/
- Object Management Group: *Business Process Model and Notation (BPMN) Version 2.0.2*, 2013. https://www.omg.org/spec/BPMN/2.0.2/