Text rein, Modell raus? Selbst nachmessen.

Begleitmaterial zur Folge 1 des Kanals «Stephan Fischli · Agentic Orchestration» · Stufen 1 bis 3 · Stand Oktober 2026

Im Video habe ich gemessen, wie gut Sprachmodelle aus einer Prozessbeschreibung ein BPMN-Modell erzeugen, und zwar an fünf Stufen, die aufeinander aufbauen. Mit diesem Paket könnt ihr die ersten drei Stufen mit eurem eigenen Werkzeug nachmessen: ChatGPT, Claude, Copilot, ein Modellierungswerkzeug mit KI oder was immer ihr nutzt. Ihr braucht dafür kein Programm, nur zehn Durchläufe, einen BPMN-Viewer und diese Prüfliste.

Im Paket

- `01_Notiz_Retourenprozess.txt`: die Eingabe für die Stufen 1–3. Eine Notiz aus dem Gedächtnis zu einem Retourenprozess, so wie solche Unterlagen in der Praxis aussehen (Englisch, genau so wurde sie im Versuch verwendet).
- `02_Aufgabe.txt`: die Bitte an die KI, unverändert aus dem Versuch.
- `03_Anleitung_und_Pruefliste.pdf`: dieses Dokument, mit Messblatt und den Vergleichswerten aus dem Video.
- `04_Messblatt.csv`: das Messblatt zum Ausfüllen in Excel.
- `…_EN`: Aufgabe, Anleitung und Messblatt auf Englisch.
- `Zusatz_Stufe4_drei_Unterlagen/`: Word-Dokument, Schulungsfolien und die Notiz. Sie widersprechen sich bewusst ein wenig; für alle, die auch Stufe 4 ausprobieren wollen.

Alle Unterlagen sind erfunden; Firmen, Personen und Abläufe sind fiktiv.

Der Versuch in drei Schritten

1 **Zehnmal fragen**
Öffnet für jeden Durchlauf eine **neue** Unterhaltung (ohne Verlauf, ohne Projektwissen) und schickt genau diese Nachricht: die Zeile aus `02_Aufgabe.txt`, eine Leerzeile, dann der ganze Text aus `01_Notiz_Retourenprozess.txt`. Keine Rückfragen beantworten, nichts nachbessern. Was nach der ersten Antwort dasteht, zählt.

2 **Ergebnis speichern**
Kopiert das XML aus der Antwort in eine Datei `lauf01.bpmn` bis `lauf10.bpmn`. Liefert das Werkzeug eine Datei zum Herunterladen, nehmt diese. Liefert es gar kein XML, ist Stufe 1 für diesen Lauf nicht erreicht.

3 **Prüfen und eintragen**
Prüft jeden Lauf nach der Prüfliste auf der nächsten Seite und tragt das Ergebnis ins Messblatt ein. Zum Schluss zählt ihr pro Stufe: wie viele von zehn Läufen sie erreichen.

> **Warum zehnmal?** Ein einzelnes Ergebnis kann gut aussehen. Ob ein Werkzeug verlässlich ist, zeigt sich erst, wenn man dieselbe Frage wiederholt. Genau das prüft Stufe 3.

Prüfliste

1 Stufe 1 · Es lädt

- Öffnet die Datei in einem BPMN-Viewer, etwa **demo.bpmn.io** (im Browser, ohne Anmeldung) oder im Camunda Modeler.
- **Erreicht**, wenn die Datei ohne Fehlermeldung öffnet **und** das Diagramm vollständig gezeichnet wird: alle Aufgaben, Entscheidungen, Ereignisse und Pfeile sind sichtbar.
- Typische Fehler aus dem Versuch: Das XML ist kaputt (z. B. ein `--` in einem Kommentar), oder der Diagramm-Teil (`bpmndi:BPMNDiagram`) fehlt ganz, dann zeigt der Viewer «no diagram to display». Beides zählt als **nicht erreicht**, auch wenn der Prozess im Text richtig beschrieben ist.

2 Stufe 2 · Vollständig

- Die Notiz beschreibt einen Normalfall und **zwei Ausnahmen**. Beide müssen als eigener Pfad vorkommen, der zu einem eigenen Ende führt:
  - **Anfrage abgelehnt**: Das Returns-Team genehmigt die Rückerstattung nicht.
  - **Ware nicht in Ordnung**: Das Lager stellt fest, dass es der falsche Artikel ist oder der Zustand nicht stimmt.
- **Erreicht**, wenn beide Ausnahmen modelliert sind. Ein Lauf, der Stufe 1 nicht erreicht, kann trotzdem auf Stufe 2 geprüft werden (im Text oder XML nachsehen); für die Gesamtwertung zählt aber nur, was auch lädt.
- Nicht verlangt: eine Frist für die Rücksendung. Die steht nur in den Zusatzunterlagen, nicht in der Notiz.

3 Stufe 3 · Stabil

- Vergleicht die zehn Läufe miteinander. Zwei Läufe haben **denselben Ablauf**, wenn sie dieselben Schritte in derselben Reihenfolge haben und dieselben Entscheidungen mit denselben Ausgängen. Unterschiede in den Namen («Prüfen» oder «Kontrollieren») und im Layout zählen **nicht**.
- Gebt jedem neuen Ablauf einen Buchstaben (A, B, C …) und tragt ihn ins Messblatt ein.
- **Ergebnis** ist die Zahl der verschiedenen Abläufe. Ideal ist 1: zehnmal gefragt, zehnmal derselbe Prozess.

> **Stufe 4 und 5 (für Neugierige).** Nehmt statt der Notiz die drei Unterlagen aus dem Zusatzordner. Die Notiz sagt: nur Rückerstattung, kein formaler Abschluss, kein Supervisor, und intern heisst es «Refund Request». Die Folien und das Word-Dokument sagen teilweise das Gegenteil.
> **Stufe 4 · Treu**: Übernimmt das Modell eine Seite still, oder erfindet es etwas? Bleibt der Begriff «Refund Request»?
> **Stufe 5 · Ehrlich**: Fragt das Werkzeug nach, bevor es entscheidet, oder markiert es seine Annahmen sichtbar im Modell?

Vergleichswerte aus dem Video

Eingabe: die Notiz · je 10 Läufe · eine Nachricht ohne Systemanweisung · Standard-Temperatur · Abruf über die Programmierschnittstelle, Oktober 2026. Ausgewertet automatisch; bei der Prüfung von Hand können einzelne Läufe anders eingestuft werden.

Werkzeug / Modell	Stufe 1 lädt mit Diagramm	Stufe 2 Pfad «Ware nicht in Ordnung»	Stufe 3 verschiedene Abläufe
gpt-oss (120b)	0 / 10	3 / 10	6
DeepSeek V4 Flash	0 / 10	1 / 10	5
GPT-4.1	0 / 10	4 / 10	7
GPT-5 mini	3 / 10	8 / 10	7
GLM-5	10 / 10	1 / 10	6
Labor mit Methode (gpt-oss)	**10 / 10**	**9–10 / 10**	**5–7**

Den Pfad «Anfrage abgelehnt» hatten alle Modelle in allen Läufen. Werte unabhängig davon, ob der Lauf auch Stufe 1 erreicht. «Labor mit Methode»: dasselbe Modell gpt-oss, eingebettet in eine Methode (festes Schema, Stilregeln aus «BPMN Method and Style», Layout-Konvention, Prüfung vor der Ausgabe); umgesetzt in meinem Labor-Werkzeug MESTRO (ELCA).

Messblatt

Werkzeug / Modell: _____ Datum: _____

Lauf	Stufe 1 lädt mit Diagramm	Stufe 2 «abgelehnt»	Stufe 2 «nicht in Ordnung»	Stufe 3 Ablauf (A, B …)	Bemerkung
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
Summe	**__ / 10**	**__ / 10**	**__ / 10**	**__ Abläufe**	

> **Eure Stufe:** Stufe 1 gilt als erreicht, wenn alle zehn Läufe laden; Stufe 2, wenn zusätzlich alle zehn beide Ausnahmen enthalten; Stufe 3, wenn zusätzlich alle zehn denselben Ablauf haben. Schreibt mir in die Kommentare unter dem Video, welches Werkzeug welche Stufe erreicht, und wo es bei euch bricht.

bpmn.ch · Stephan Fischli · Begleitmaterial zur Folge 1 «Text rein, Modell raus?». Frei verwendbar für eigene Messungen und Schulungen, mit Quellenangabe.