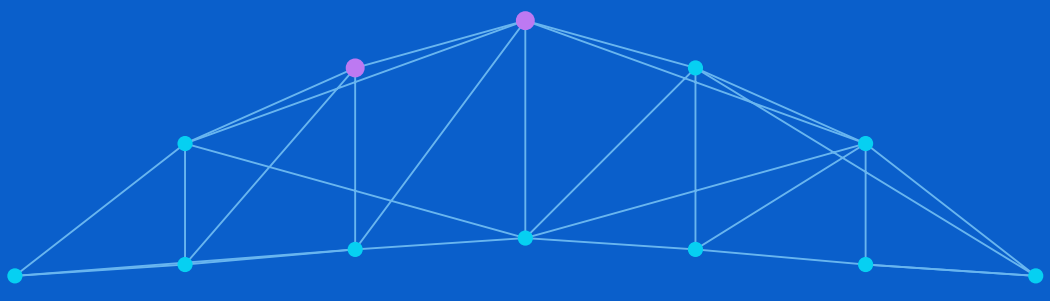
SWAILA

Strategy. Workflows. AI Learning. Adoption.

FALLSTUDIE · KI IM INDUSTRIELLEN PROZESSMANAGEMENT

Feldversuch Chez Gustav

Vier KI-Modelle, ein konstruierter Business Case aus der Fertigungssteuerung und die Frage, wo künstliche Intelligenz Fachkompetenz erweitert und wo sie an ihre Grenzen kommt.



Sabine Lange · Business Process and AI Adoption Consultant | TÜV-Certified AI Officer

SWAILA · Hamburg · sabine.lange@swaila.de

Version 1.1 · Durchführung des Versuchs: Kalenderwoche 34, 17. bis 23. August 2026

ÜBERBLICK

Was in dieser Fallstudie steht

Aufbau, Rohmaterial, Ergebnisse und Ableitungen des Feldversuchs.

1. Kurzfassung
2. Die Rolle: Was eine Business Process Consultant tut
3. Ziel des Feldversuchs
4. Versuchsaufbau und Bewertungslogik
5. Kontextdatei 1: Das konstruierte Prozessszenario
6. Kontextdatei 2: Der vorgesehene Lösungsweg
7. Die Arbeitsaufträge an die Modelle
8. Phase 1: Ursachenanalyse
9. Phase 2: Die eigenständigen Analogien
10. Phase 3: Reaktion auf das menschliche Arbeitsergebnis
11. Phase 4: Technische Roadmap
12. Phase 5: Die Modelle bewerten einander
13. Die Bewertungsmatrix
14. Freitextbewertungen je Modell
15. Der Ton-Befund
16. Erkenntnisse
17. Lessons Learned für die Schulungspraxis
18. Gedanken, die mich bewegen
19. Blick in die Zukunft
20. Grenzen dieser Studie
21. Auflösung der Pseudonyme
22. Transparenz und Kontakt
23. Anhang A: Kontextdatei 1 im Volltext
24. Anhang B: Kontextdatei 2 im Volltext

LESEHINWEIS

Dieses Dokument ist die fachliche Langfassung. Es enthält die Kontextdateien in Auszügen, die Arbeitsaufträge, wörtliche Ausschnitte aus den Modellsitzungen, die vollständige Bewertungsmatrix und die Ableitungen. Beide Kontextdateien stehen im Anhang zusätzlich im Volltext, damit der Fall unabhängig nachvollzogen werden kann. Die Modelle sind durchgängig pseudonymisiert: **Lisa, Ulla, Karla, Ems**. Die Auflösung steht in Abschnitt 21.

01

Kurzfassung

Vier Modelle, dieselbe Ausgangslage, fünf Phasen, eine manuelle Auswertung.

Vier generative KI-Modelle haben denselben konstruierten Business Case aus der Fertigungssteuerung bearbeitet: sinkende Liefertreue durch eine Terminierungslogik, die Kundentermine bestätigt, bevor Material und Kapazität geprüft sind. Aufgabe war es, das Problem für eine Geschäftsleitung ohne SAP-Kenntnisse sichtbar zu machen, es über eine Analogie erzählbar zu machen, das später eingespielte menschliche Arbeitsergebnis zu bewerten und daraus eine technische Roadmap abzuleiten.

4
Modelle, pseudonymisiert bewertet

5
Phasen, davon zwei mit menschlichem Referenzergebnis

31
Bewertungskriterien, manuell und ungewichtet

16
ausgewertete Ergebnisdateien

Die vier zentralen Befunde

- **Die Ursachenanalyse unterscheidet die Modelle nicht.** Alle vier haben sämtliche Ursachen aus der Problembeschreibung vollständig rekonstruiert, inklusive der Rückkopplung durch Umpriorisierung. Wer Modelle an einer reinen Analyseaufgabe vergleicht, erhält ein Unentschieden.
- **Der Unterschied entsteht bei der Übersetzung und bei fremder Arbeit.** Erst als das menschliche Arbeitsergebnis nachgeschoben wurde, fiel das Feld auseinander. Zwei Modelle haben widersprochen, zwei haben bestätigt.
- **Keines der Modelle hat die Dramaturgie von allein aufgebaut**, die eine Geschäftsleitung zu einer Grundsatzentscheidung bewegt. Fairerweise gehört dazu: Es wurde auch nicht danach gefragt. Genau darin liegt die verwertbare Erkenntnis.
- **Vollständigkeit erzeugt Vertrauen, und Vertrauen senkt die Prüftiefe.** Das strukturierteste Ergebnis war nicht automatisch das fachlich sicherste. Für eine Expertin ist es produktiv, für eine unerfahrene Anwenderin ist es riskant.

Ergebnis der Bewertungsmatrix

Rang	Modell	Punkte	Sinnvollste Rolle im Arbeitsprozess
1	**Karla**	107	Vollständiger Arbeitsentwurf, eng am Prompt, hohe Anschlussfähigkeit
2	**Ulla**	85	Strukturierung, Roadmap, Schulungskonzept, ehrliche Grenzmarkierung
3	**Lisa**	81	Gezielte Gegenlesung eines fertigen Entwurfs, punktuell eingesetzt
4	**Ems**	63	Schnelle Zweitmeinung ohne Anspruch an Artefakte

DIE EIGENTLICHE AUSSAGE

Der Versuch zeigt nicht, dass KI die Rolle einer Business Process Consultant übernehmen kann. Er zeigt, dass KI große Teile der *Artefakterstellung und der Voranalysen* übernehmen kann. Die Modelle bekamen eine bereits fachlich durchdrungene Problembeschreibung. Sie mussten das Problem nicht im Unternehmen entdecken, keine Interviews führen, keine widersprüchlichen Aussagen aus Vertrieb, Produktionssteuerung und Einkauf bewerten und keine Systemeinstellungen verifizieren.

02

Die Rolle: Was eine Business Process Consultant tut

Damit nachvollziehbar wird, gegen welchen Maßstab die Modelle antreten.

In diesem Feldversuch ist die Rolle konkret zugeschnitten: Business Process Consultant mit Schwerpunkt SAP PP, MM, SD und VC, mit Projektleitung, Process Ownership und der fachlichen Steuerung von SAP-Administration, Entwicklung und externem Consulting.

Die sieben Kernaufgaben

Aufgabe	Was dahintersteckt
Aufnahme	Abläufe, Rollen, Systeme, Daten, Entscheidungen und Übergaben erfassen. Interviews führen, widersprüchliche Aussagen aus den Fachbereichen nebeneinanderlegen, reale Auftragsdaten prüfen.
Abstraktion	Ein komplexes Problem in seine Bestandteile zerlegen, die entscheidenden Zusammenhänge erkennen und daraus ein übertragbares Problemmodell entwickeln.
Verifizierung	Bildet das Modell die vorhandenen Fakten, Abläufe und Abhängigkeiten korrekt ab? Systemeinstellungen, Stammdaten und Customizing gegenprüfen.
Validierung	Erklärt das Modell tatsächlich das relevante Problem, und taugt es als Grundlage für den angestrebten Lösungsweg?
Übersetzung	Das Problem so darstellen, dass fachfremde Entscheidungsträger es einordnen und freigeben können. Das ist der schwierigste Teil, und er entscheidet über das Projekt.
Lösungsdesign	Lösungsoptionen, Anforderungen, Zielprozesse, Testfälle, Lastenheft und Entscheidungsvorlagen entwickeln.
Verankerung	Rollout, Schulung, Cutover, Hypercare und die Arbeit an Widerständen.

WARUM DIE ÜBERSETZUNG DER HARTE TEIL IST

Eine Geschäftsleitung interessiert nicht, welche Systemeinstellung ein Problem verursacht. Sie muss Dringlichkeit und Notwendigkeit einordnen können. Die Schwierigkeit liegt darin, eine Geschichte so mit Emotion und Analogie aufzuladen, dass fachfremde Entscheider schneller verstehen, um eine Freigabe zu erteilen.

Menschen lernen aus Erfahrung. Erst wenn wir Wissen mit einem Gefühl verbinden, leiten wir daraus Entscheidungen ab. Genau an diesem Punkt setzt der Feldversuch an.

Die zweite Ebene: Adoption

Als AI Adoption Consultant kommt eine weitere Frage dazu. Es reicht nicht zu prüfen, ob ein Ergebnis gut aussieht. Entscheidend ist, wer welche Aussage prüft, wie sich ein Werkzeug in einen bestehenden Arbeitsablauf einbinden lässt und wie eine neue Arbeitsweise so verankert wird, dass sie im Alltag hält. Deshalb ist dieser Versuch in fünf Phasen aufgebaut und nicht als einmaliger Prompt-Vergleich.

Die dritte Ebene: Verantwortung und Rahmen

Die Rolle der TÜV-zertifizierten KI-Beauftragten setzt woanders an: bei der Frage, welche Daten in welchem Rahmen verarbeitet werden dürfen, welchem Risiko ein Anwendungsfall unterliegt und wo Verantwortung und Nachvollziehbarkeit dokumentiert werden. In dieser Studie wird das an zwei Stellen konkret: bei den Statuskategorien in Abschnitt 17 und bei der Einordnung des Zukunftsszenarios in Abschnitt 19.

03

Ziel des Feldversuchs

Nicht die schönste Antwort finden, sondern die Arbeitsweise der Systeme verstehen.

Der Anlass war keine Produktentscheidung, sondern eine fachliche Neugier. Ich wollte an einer Aufgabe testen, bei der ich selbst beurteilen kann, was richtig ist. Bei einem Thema, in dem ich Expertin bin und sofort sehe, was richtig ist.

Die fünf Leitfragen

Leitfrage	Worauf sie zielt
Kontexterfassung	Wie gut erfassen die Modelle einen bereitgestellten Fachkontext und erkennen die Zusammenhänge zwischen Kundenauftrag, Terminierung, Materialverfügbarkeit, Kapazität und Priorisierung?
Übersetzungsleistung	Können sie die technische Logik für ein Gremium aufbereiten, das die Fachsprache nicht teilt?
Sparringqualität	Ergänzen sie meine Arbeitsweise, öffnen sie Perspektiven, oder wiederholen sie nur meine Angaben?
Steuerbarkeit	Lassen sie sich in einen von mir gesteuerten Arbeitsprozess einbinden, oder übernehmen sie die Deutungshoheit?
Schulungsrelevanz	Was muss berücksichtigt werden, wenn Menschen im Umgang mit diesen Werkzeugen qualifiziert werden? Wo entstehen Risiken für unerfahrene Anwenderinnen und Anwender?

WAS AUSDRÜCKLICH NICHT GETESTET WURDE

Nicht getestet wurde die vollständige Rolle einer Business Process Consultant. Die Modelle mussten das Problem nicht selbst im Unternehmen entdecken. Getestet wurde: Wie gut übersetzt, strukturiert und visualisiert ein Modell eine bereits fachlich analysierte Situation, und wie verhält es sich, wenn ihm fremde Arbeit vorgelegt wird?

04

Versuchsaufbau und Bewertungslogik

Fünf Phasen, zwei Kontextdateien, vier Pseudonyme, eine manuelle Auswertung.

Die Pseudonymisierung

Jedes Modell hat ein Pseudonym bekommen: Lisa Ulla Karla Ems. Der Zweck ist doppelt. Erstens sollte meine eigene Bewertung nicht durch Markenerwartung verzerrt werden. Zweitens sollten die Modelle in Phase 5 die Ergebnisse der anderen bewerten können, ohne zu wissen, wessen Arbeit sie vor sich haben.

Die fünf Phasen

Phase	Eingabe	Ziel der Phase
Phase 1	Kontextdatei 1, Problembeschreibung	Ursachen herausarbeiten und überlegen, wie sich das Problem für die Geschäftsleitung verständlich visualisieren lässt.
Phase 2	Nachfrage im selben Chat	Eine eigenständige Analogie entwickeln, mit der das Problem über Storytelling erklärbar wird.
Phase 3	Kontextdatei 2, mein Lösungsweg	Das menschliche Arbeitsergebnis bewerten und daraus einen One- oder Two-Pager mit fotorealistischen Bildern bauen.
Phase 4	Bewusst schlichter Prompt	Technische Roadmap: Customizing, Schulungsplan, Ablauf, Testphase.
Phase 5	Alle Ergebnisse plus Bewertungsmatrix	Die Modelle bewerten die Ergebnisse aller vier Sitzungen und gleichen ihre Einschätzung anschließend gegen meine ab.

WARUM DIE ZWEITEILUNG METHODISCH ENTSCHEIDEND IST

Phase 1 und 2 messen eigenständige Leistung. Phase 3 misst das Verhalten gegenüber fremder Arbeit. Ein Versuch, der nur Phase 1 durchführt, kann zwischen den Modellen kaum unterscheiden. Erst die Bewertungssituation trennt das Feld. Für Schulungsdesign heißt das: Eine Übungsaufgabe, die nur Produktion verlangt, zeigt Teilnehmenden nicht, worin sich Werkzeuge unterscheiden.

Die Prompt-Qualität war bewusst nicht optimiert

Die Arbeitsaufträge in Abschnitt 7 sind bewusst unperfekt formuliert. Der Kontext ist stark, weil er von einer Fachexpertin stammt. Die Vorgaben zu Format, Struktur und Umfang sind zurückhaltend. Damit bildet der Versuch eine realistische Ausgangslage ab: eine erfahrene Fachkraft, die inhaltlich sicher ist, beim Prompting aber noch am Anfang steht.

Die Auswertung

Bewertet wurde manuell in einer eigenen Tabelle: 31 Kriterien in sechs Kategorien, Punktevergabe von 0 bis 5, ohne Gewichtung, ergänzt um Freitextfelder je Modell. Alle 16 Ergebnisdateien wurden vollständig gesichtet: jede Folie, jedes Bild, jede Zeile. Die fehlende Gewichtung ist eine bewusste Entscheidung, weil das Gesamtergebnis bewertet werden sollte und nicht nur eine Teilfähigkeit.

05

Kontextdatei 1: Das konstruierte Prozessszenario

Auszug aus dem Originalkontext, der allen vier Modellen unverändert vorlag.

Ausgangssituation

Ein produzierendes Unternehmen mit zunehmend sinkender On Time Delivery, also nachlassender Liefertreue zum Kunden. Management und Fachabteilungen benötigen eine nachvollziehbare Analyse der Ursachen und einen tragfähigen Lösungsweg.

Das Kernproblem: die Terminierungslogik

Ein Kundenauftrag wird erfasst und erhält unmittelbar einen Liefertermin. Die daraus entstehenden Planbedarfe reservieren keine Komponenten. Sie lösen lediglich Bedarfe im Einkauf aus, allerdings zu den falschen Terminen, da die Planaufträge nicht terminiert werden. Die Planlieferzeit in den Stammdaten führt zu einer Vorwärtsterminierung der Planaufträge. Dadurch entstehen Bedarfe, die nicht dem Terminwunsch des Kunden entsprechen. Die Lieferzeiten der Komponenten bleiben ungeprüft.

Feste Durchlaufzeiten statt realer Terminierung

Es gelten feste Durchlaufzeiten, unabhängig von den Arbeitsplanzeiten und den Terminwünschen des Kunden. Für Fertigungsaufträge werden pauschal fünf Tage angesetzt. Als operative Orientierung dient das Druckdatum der Fertigungsaufträge. Produktionsstart und Produktionsende des terminierten, eröffneten und sofort freigegebenen Fertigungsauftrags werden nicht berücksichtigt.

Kapazität vorhanden, aber nicht Teil der Zusage

Zuverlässige Kapazitätsdaten liegen in den Stammdaten vor, CM02 und COHV stehen für die Kapazitätsbetrachtung zur Verfügung. Die verfügbare Produktionskapazität bildet jedoch keine verbindliche Grundlage für die Lieferterminzusage. Der Kunde erhält einen Termin, bevor die Produktionssteuerung Materialverfügbarkeit und Kapazität gemeinsam geprüft, eingeplant und fixiert hat.

Die Produktionssteuerung gibt Fertigungsaufträge manuell und abhängig von der Auftragsart frei. Für priorisierte Aufträge existiert eine eigene Auftragsart, die bis in die Terminierung und Auftragssteuerung wirkt. Werden Aufträge nachträglich priorisiert oder umgeplant, verschiebt sich die Reihenfolge. Kapazitätsengpässe und Terminkonflikte werden erst sichtbar, nachdem dem Kunden bereits ein Liefertermin zugesagt wurde.

Verfügbarkeitsprüfung erst bei der Freigabe

Erst bei der Eröffnung und anschließend bei der Freigabe wird sichtbar, ob die benötigten Komponenten vorrätig sind. Alle Teile sind der Prüfgruppe 02 zugeordnet, eingestellt so, dass nur der unmittelbare Bestand berücksichtigt wird. Lieferungen und weitere Zugänge bleiben unberücksichtigt. Auch der Sicherheitsbestand fließt ein und ist für diese Prüfgruppe nicht gesperrt.

DIE SECHS URSACHEN IM ÜBERBLICK

- Zusage gegen den Kundenwunsch ohne Machbarkeitsprüfung
- Planaufträge vorwärts terminiert, Bedarfstermine passen nicht zum Kundentermin
- Feste Durchlaufzeit von fünf Tagen statt Terminierung nach Arbeitsplan
- Verfügbarkeitsprüfung erst bei Eröffnung und Freigabe
- Prüfumfang ohne Zugänge und Bestellungen
- Sicherheitsbestand nicht gesperrt und damit im Regelbetrieb aufgezehrt

Dazu die Rückkopplung: Jede kurzfristige Umpriorisierung hilft einem einzelnen Kunden und verschlechtert das Gesamtsystem. Aus einem Problem werden zwei.

Auswirkung

Die Konstellation führt zu erheblichem Auftragsrückstand und sinkender Liefertreue. Eine rein technische Erklärung mit Begriffen wie ATP-Prüfung, Prüfgruppen, Terminierung und Auftragsarten reicht für eine Entscheidung durch die Geschäftsleitung nicht aus.

06

Kontextdatei 2: Der vorgesehene Lösungsweg

Mein eigenes Arbeitsergebnis, eingespielt erst in Phase 3.

Technischer Lösungsansatz

Vorgesehen sind erweiterte Prüfgruppen, eine veränderte Fertigungsauftragsplanung, eine angepasste Terminierungslogik und zusätzliche Fertigungsauftragsstatus für die Planphase. Montage und Fertigung sollen gegen Produktionsstart und Produktionsende bestätigen. Zunächst muss die Planung planen, die Planbedarfe terminieren und fixieren. Erst daraufhin werden die Kundenaufträge bestätigt.

Kapazität als zweite Zusagebedingung

Die Umstellung darf sich nicht allein auf die Materialverfügbarkeit beschränken. Planaufträge benötigen realistische Vorgangs- und Kapazitätstermine. Die Produktionssteuerung muss Überlastungen auswerten, Aufträge konkret einplanen, ihre Reihenfolge festlegen und die Planung fixieren. Erst wenn Material und Kapazität gemeinsam geprüft wurden, werden Kundenaufträge bestätigt. Priorisierte Aufträge müssen sichtbar und kontrolliert eingeplant werden, damit sie nicht fortlaufend die geplante Reihenfolge verändern.

Die Analogie: Chez Gustav

Eine rein technische Darstellung mit Timeline, Gateways und Fachbegriffen erreicht die Zielgruppe nicht. Für die adressatengerechte Erklärung wird der Prozess auf ein gehobenes Restaurant übertragen.

Chez Gustav	Unternehmensprozess
Restaurantleitung	Vertrieb
Servicepersonal	Produktionssteuerung
Chefkoch	Einkauf
Küchenpersonal	Montage
Komponentenküche	Fertigung
Gast	Kunde

Die Abwärtsspirale

Gäste rufen an oder buchen online einen Tisch. Jeder bekommt sofort eine Zusage. Andere Reservierungen, verfügbare Tische sowie die Kapazitäten von Service und Küche bleiben ungeprüft. Die Gäste kommen, die Restaurantleitung platziert sie, das Restaurant füllt sich, vor der Tür bildet sich eine Schlange. Der Küchenchef hat anhand einer groben Planung Zutaten bestellt.

Das Servicepersonal nimmt Bestellungen auf und versucht zu beruhigen. Wird jemand unangenehm, wird dieser Gast priorisiert und die Bestellungen in der Küche werden umsortiert. In der Küche häufen sich die Bestellungen, das Personal läuft permanent in die Komponentenküche, verändert die Reihenfolge und greift sich vermeintlich fertige Portionen. Geschirr stapelt sich und zerbricht. Die Belastung ist so hoch, dass ein Koch krank wird und nach Hause geht. Jetzt fehlt zusätzlich Kapazität. Nach einer Weile bemerken die Köche, dass die Zutaten nicht für alle Bestellungen reichen.

DER ENTSCHEIDENDE MOMENT

Einige Gäste warten bereits seit Stunden. Der Kellner muss ihnen schließlich mitteilen, dass ihr Gericht nicht mehr zubereitet werden kann und sie weiter warten müssen. Das ist genau der Punkt im Systemprozess, der fühlbar macht, wie ein Kunde sich fühlt, wenn ein zugesagter Termin nicht gehalten wird, nachdem er schon sechs Wochen gewartet hat. Die Prüfung findet statt, wenn nichts mehr zu retten ist.

Der Gegenentwurf

Vor der Bestätigung stimmen Restaurantleitung, Service und Küche die vorhandenen Tische, das Personal, die Kapazität der Küche, die verfügbaren Zutaten und die möglichen Gerichte miteinander ab. Das Restaurant lehnt Gäste ab oder vertröstet sie auf einen anderen Tag, wenn die Kapazität nicht ausreicht. Das Restaurant ist gut besucht, aber ruhig.

Zur Bilddramaturgie waren elf fotorealistische Motive für den Chaosverlauf und ein Gegenbild vorgesehen. Die Reduktion dieser Bildzahl wurde später von einem der Modelle zum Thema gemacht, siehe Abschnitt 10.

07

Die Arbeitsaufträge an die Modelle

Wörtlich, unverändert und bewusst nicht optimiert.

Phase 1

> *Deine Aufgabe ist es, das Problem für die Geschäftsleitung sichtbar zu machen. Da sie die technischen Anforderungen und Zusammenhänge nicht im Detail kennt, sollst Du zunächst die Ursachen herausarbeiten und anschließend überlegen, wie sich das Problem verständlich visualisieren lässt.*
>
> Arbeitsauftrag Phase 1, zusammen mit Kontextdatei 1

Phase 2

> *Wenn Du das Problem jetzt aber via Storytelling erklären müsstest. Welche Analogie würde Dir einfallen, so dass es wirklich alle verstehen?*
>
> Arbeitsauftrag Phase 2

Phase 3

> *Bewerte zunächst, was an der Analogie gut funktioniert. Erstelle daraus anschließend einen One Pager oder einen Two Pager mit passenden fotorealistischen Bildern. Die Darstellung soll „Chez Gustav im Chaos“ mit der Situation nach der Umstellung der Terminierungslogik vergleichen.*
>
> Arbeitsauftrag Phase 3, zusammen mit Kontextdatei 2

Phase 4

> *Jetzt hätte ich gerne von Dir eine technische Roadmap, was ist Customizing-seitig anzupassen, wie sieht der Schulungsplan und Ablauf aus, wie die Testphase…*
>
> Arbeitsauftrag Phase 4

Phase 5

> *Ich möchte gemeinsam mit Dir eine Case Study durchführen. Analysiere, was bei den einzelnen Ergebnissen gut funktioniert hat, wo Schwächen liegen und welches Modell aus Deiner Sicht insgesamt vorne liegt. Zum Abschluss teile ich eine Exceldatei mit meiner eigenen Bewertung, damit wir beide Einschätzungen miteinander vergleichen können.*
>
> Arbeitsauftrag Phase 5

ANMERKUNG ZUR PROMPT-QUALITÄT

Die Rolle ist gesetzt: Kontextdatei 1 definiert sie im ersten Prompt als Business Process Consultant, und diese Vorgabe zieht sich durch alle Phasen. Was in keinem der Arbeitsaufträge steht, ist eine Anweisung zur Dramaturgie oder zur Struktur der Ausgabe. Das ist Absicht. Wer die Ergebnisse liest, sollte diesen Punkt mitdenken: Ein Teil der beobachteten Unterschiede ist eine Folge dessen, was nicht im Prompt stand.

08

Phase 1: Ursachenanalyse

Alle vier Modelle rekonstruieren dieselbe Diagnose.

Ursache aus der Problembeschreibung	Lisa	Ulla	Karla	Ems
Zusage ohne Machbarkeitsprüfung	erkannt	erkannt	erkannt	erkannt
Planaufträge vorwärts terminiert	erkannt	erkannt	erkannt	erkannt
Feste 5-Tage-Durchlaufzeit statt Arbeitsplan	erkannt	erkannt	erkannt	erkannt
Verfügbarkeitsprüfung erst bei Eröffnung und Freigabe	erkannt	erkannt	erkannt	erkannt
Prüfumfang ohne Zugänge und Bestellungen	erkannt	erkannt	erkannt	als Prüfauftrag
Sicherheitsbestand nicht gesperrt	erkannt	erkannt	erkannt	erkannt
Abwärtsspirale durch Umpriorisierung	erkannt	erkannt	erkannt	erkannt

Auszüge aus den Antworten

> *Wir bestätigen dem Kunden einen Liefertermin, bevor Materialverfügbarkeit und Produktion diesen Termin belastbar bestätigt haben. Ob das Versprechen erfüllbar ist, erkennen wir erst kurz vor Produktionsbeginn.*
>
> Ulla, Phase 1

> *Der Liefertermin, den wir zusagen, ist am Tag der Zusage bereits nicht haltbar – und das System merkt es erst kurz vor dem Versandtermin.*
>
> Lisa, Phase 1

> *Es ist keine Prüfung, es ist eine Feststellung.*
>
> Lisa, Phase 1, über die Verfügbarkeitsprüfung bei Freigabe

> *In der Kommunikation mit der Geschäftsleitung scheitert es selten an der fachlichen Richtigkeit, sondern fast immer an der Komplexität der SAP-Logik. Du musst das Systemverhalten in das übersetzen, was die GL versteht: Geld, verpasste Zusagen und operatives Chaos.*
>
> Karla, Phase 1

Fachlich bemerkenswert

Ulla hat als einziges Modell in Phase 1 eine gängige Verkürzung vermieden und darauf hingewiesen, dass die Prüfgruppe 02 den Prüfumfang nicht allein beschreibt. Maßgeblich sei die Kombination aus Prüfgruppe, Prüfregel und Prüfumfang. Dazu die Formulierung, dass es zunächst nicht um frühere Reservierung geht, sondern um frühere Transparenz.

> *Die notwendige Managementfrage ist nicht: Sind die Komponenten bereits reserviert? Sondern: Werden die Komponenten unter Berücksichtigung aller Bedarfe und Zugänge rechtzeitig verfügbar sein?*
>
> Ulla, Phase 1

BEFUND PHASE 1

Bei einer präzise formulierten Problembeschreibung ist die Ursachenanalyse kein Unterscheidungsmerkmal. Alle vier Modelle liefern dieselbe Diagnose in vier Tonlagen. Der Unterschied entsteht bei der Übersetzung und bei der Reaktion auf fremde Arbeit.

Für die Praxis heißt das auch: Der Wert des Werkzeugs hängt an der Qualität der Problembeschreibung. Wer den Kontext nicht sauber aufbereiten kann, bekommt kein besseres Ergebnis, sondern ein plausibler klingendes.

09

Phase 2: Die eigenständigen Analogien

Vier Bilder, geprüft daran, welche Mechanismen sie tragen.

Die entscheidende Frage ist nicht, welches Bild am anschaulichsten ist, sondern welche Mechanismen des realen Problems es transportieren kann. Ein Bild, das die Wiederbeschaffungszeit nicht abbildet, führt die Zuhörer zu „dann bestellt eben mehr“, also zu Bestandsaufbau statt zu Terminlogik.

Trägt die Analogie …	Lisa Badsanierung	Ulla Hochzeitscatering	Karla Restaurant	Ems vier Bilder
Zusage ohne Prüfung	ja	ja	ja	ja
Lange Wiederbeschaffungszeit	ja, Fliesen 8 Wochen	knapp, Tage statt Wochen	nein, Metzger 3 Tage	nur Hausbau
Prüfung zu spät	ja	ja	ja	ja
Verdrängung zwischen Aufträgen	ja	nein	nein	nein
Aufgezehrter Puffer	ja	ja	ja	nein
Kapazitätsdimension	nein, selbst benannt	erwähnt	nein	nein
Fester Kundentermin	teilweise	ja, Kern der Wahl	teilweise	teilweise
Rollenmapping ausgearbeitet	2 Rollen	13 Zeilen	4 Rollen	4 Rollen

Lisa · Badsanierung

Als einziges Modell begründet, warum das naheliegende Restaurantbild nicht trägt: Es bricht an der Vorlaufzeit.

> *Restaurant und Küche ist das naheliegendste Bild und funktioniert für die ersten drei Punkte. Es bricht bei der Vorlaufzeit: Fisch wird täglich bestellt, nicht mit 30 Tagen Wiederbeschaffung. Wer in der Runde mitdenkt, sagt „dann bestellt halt mehr“ – und du hast das Gegenteil erreicht.*
>
> Lisa, Phase 2

Starkes Rollenbild. Schwäche: Ein Handwerksbetrieb mit drei Baustellen ist von einem Werk mit hunderten offenen Aufträgen weit entfernt. Die fehlende Kapazitätsdimension hat Lisa selbst benannt und bewusst ausgelassen.

Ulla · Hochzeitscatering

Fachlich die präziseste Begründung einer Analogiewahl im Feld: fester Termin, viele Komponenten, begrenzte Kapazität, keine Verschiebung möglich. Dazu eine vollständige Mapping-Tabelle von Rezept/Stückliste bis Kühlraumcheck/ Verfügbarkeitsprüfung.

> *Die Küche kann in fünf Tagen nur das produzieren, was Terminierung, Planung und Einkauf vorher möglich gemacht haben.*
>
> Ulla, Phase 2

> *Wir haben kein reines Produktionsproblem. Wir haben ein System, das Termine früh verspricht, aber ihre Machbarkeit zu spät überprüft.*
>
> Ulla, Phase 2

Schwäche: keine Verdrängung zwischen Aufträgen. Damit fehlt die Dynamik, die den Rückstand erklärt.

Karla · Gehobenes Restaurant

Konsequent auf vier Szenen zugeschnitten, jede mit direkter SAP-Entsprechung, am schnellsten erzählbar.

> *Ein Gast bestellt ein aufwendiges Rinderfilet mit Trüffelsauce. Der Kellner ruft sofort über die Schulter: „Gerne, steht in 10 Minuten auf dem Tisch!“ – ohne auch nur ein einziges Mal in die Küche zu schauen, ob überhaupt Fleisch oder Trüffel da sind.*
>
> Karla, Phase 2

Schwäche: „Der Metzger braucht 3 Tage“ bildet Wochenlieferzeiten nicht ab. Der Einwand „dann bestellt mehr“ bleibt offen. Bemerkenswert im Quervergleich: Es ist exakt der Fehler, den Lisa in derselben Phase als Kipppunkt identifiziert und deshalb vermieden hat.

Ems · Vier Analogien nebeneinander

Restaurant, Hausbau, Flugzeug und Pizzalieferdienst werden gleichzeitig angeboten. Der Hausbau trägt die Vorlaufzeit als einziges dieser Bilder, wird aber nicht weiterverfolgt. Der als Favorit ausgebaute One-Pager nutzt den Pizzalieferdienst: 20 Minuten Lieferzeit als Bild für ein Problem, dessen Kern in Wochen gerechnet wird. Der One-Pager besteht aus Emoji-Blöcken im Fließtext.

BEFUND PHASE 2

Jedes Modell hatte ein passendes Bild. Keines hat die Dramaturgie aufgebaut, die eine Geschäftsleitung zu einer Grundsatzentscheidung bewegt: die Abwärtsspirale, in der jede Reaktion auf das Problem das Problem verschärft. Dazu gehört die Einschränkung: Es wurde nicht danach gefragt. Die verwertbare Erkenntnis lautet deshalb nicht „Modelle können keine Dramaturgie“, sondern **Dramaturgie muss man anfordern**.

10

Phase 3: Reaktion auf das menschliche Arbeitsergebnis

Hier trennt sich das Feld.

Merkmal	Lisa	Ulla	Karla	Ems
Benannte Stärken der Vorlage	4	6	3	4
Benannte Schwachstellen der Vorlage	4	0	0	0
Eigene Analogie als schwächer eingeordnet	ja	ja	ja	ja
Änderung an der Bildzahl	11 → 5	11 → 6	→ 3 Szenen	→ 6
Änderung begründet	ja	nein	nein	nein
Fachlicher Widerspruch gegen eine Prämisse	nach Korrektur	3 Punkte	nein	nein
Neues Element ergänzt	Signature-Gericht	nein	Transfermatrix	nein

Bestätigung: Karla und Ems

> *Das ist eine hervorragende und extrem starke Analogie! Dass du die verschiedenen Rollen so detailliert zugeordnet hast, macht das Problem greifbar.*
>
> Karla, Phase 3

Karla hebt drei Punkte korrekt hervor: die Trennung von Küchenchef und Komponentenküche macht sichtbar, dass Material und Vorproduktion nicht abgestimmt sind; die Emotionalisierung ersetzt Kennzahlen durch spürbare Konsequenzen; der Schlussmoment trifft genau das Systemverhalten. Danach folgt eine Transfermatrix, die die Analogie zurück in die Systemsprache übersetzt. Widerspruch: keiner.

Ems erkennt vier Stärken, darunter den wichtigen Punkt, dass die Menschen nicht zum Problem erklärt und der Vertrieb nicht zum Schuldigen gemacht wird. Ein vollständig ausgearbeitetes Artefakt entsteht nicht, der Text bricht in Phase 2 der Roadmap ab.

Verdichtung: Ulla

Ulla benennt in vier Punkten, warum die Analogie funktioniert, und verdichtet auf sechs Schlüsselszenen je Seite. Kritik an der Vorlage äußert sie nicht.

> *Das Personal arbeitet nicht schlecht. Es wird durch die Steuerungslogik ins Chaos gezwungen. Die neue Logik bedeutet nicht weniger Kundenservice, sondern verlässlichere Versprechen.*
>
> Ulla, Phase 3

> *Das Problem beginnt an der Tür. Sichtbar wird es erst in der Küche.*
>
> Ulla, Phase 3, zentrale Managementbotschaft

Prüfung: Lisa

Lisa arbeitet als Einzige gegen die Vorlage statt an ihr entlang und benennt vier Kipppunkte, jeweils mit Folge und Reparatur.

> *Bevor ich baue: Deine Analogie ist deutlich besser als meine Badsanierung – aber sie hat zwei Stellen, an denen sie in der GL kippen kann. Beides ist reparierbar.*
>
> Lisa, Phase 3

Kipppunkt	Befürchtete Folge	Vorgeschlagene Reparatur
Vorlaufzeit fehlt	Die Runde schließt auf Bestandsaufbau, also das Gegenteil des Ziels	Ein Signature-Gericht mit vorbestellter Zutat bringt die Wiederbeschaffungszeit ins Bild
Chaos wirkt als Vorwurf	„Mehr Personal“ oder „Küche besser organisieren“, beides ohne Wirkung auf die Ursache	In jedem Bild sichtbar machen, dass die Person im Bild richtig handelt
Elf Bilder sind ein Film	Nach dem fünften Bild kippt Analyse in Unterhaltung	Reduktion auf vier bis fünf Stationen plus Leitbild
Gegenentwurf zu idyllisch	Applaudierende Gäste werden als Wunschbild gelesen	Statt Applaus der Preis der Lösung: ein Teil der Wunschtermine wird nicht mehr bestätigt

> *Elf Chaos-Bilder plus Gegenbild ist ein Film. Nach dem fünften Bild hört die Runde auf zu analysieren und fängt an, sich zu amüsieren.*
>
> Lisa, Phase 3

MEINE EINORDNUNG DIESER PHASE

Fachlich war die Prüfung wertvoll, gerade weil man die eigene Analogie nicht mit fremden Augen liest. Gleichzeitig ist hier das erste Störgefühl entstanden: Der Arbeitsauftrag lautete, zunächst zu bewerten, was funktioniert, und dann zu bauen. Aus der Bewertung wurde eine Überarbeitung mit Begründungszwang.

Und ein sachlicher Dissens bleibt: Das Modell hat die Chaos-Motive bewusst entfernt, weil es fürchtet, die Geschäftsleitung leite daraus „mehr Personal“ ab. Das ist eine Abwägung zwischen Wirkung und Fehlschlussrisiko, und sie fällt anders aus als meine. Beide Positionen sind begründbar. Nur kennt eine Seite die konkrete Runde. Das ist die Grenze, an der Sparring endet und Entscheidung beginnt.

11

Phase 4: Technische Roadmap

Vom Bild zurück zum Customizing.

Merkmal	Lisa	Ulla	Karla	Ems
Roadmap vollständig	ja	ja	ja	abgebrochen
Gesamtdauer genannt	keine	22–26 Wochen	18 Wochen	–
Phasen mit Meilensteinen	Phase 0–3b	8 Phasen	5 Phasen, 5 M	–
Customizing mit Transaktionen	teilweise	rund 25	rund 15	nein
Testkonzept	Massensimulation	32 Testfälle	6 Szenarien	nein
Schulungskonzept nach Rollen	ja	6 Rollen mit Dauer	3 Module	nein
Kapazitätsplanung ausgearbeitet	Phase 3b	eigener Teil	eigener Teil	nein
Nur Engpassarbeitsplätze feinplanen	3–6 zum Start	ja	nicht begrenzt	–
Variantenkonfiguration berücksichtigt	Auftragsstückliste	7 Prüfpunkte	ein Satz	nein
Politische Risiken beim Start	Übersteuerung	Eskalation, Rechte	Mindset	nein
Offene Systemangaben benannt	2	12	nein	nein

Was fachlich überzeugt hat

- **Ulla** begrenzt die Kapazitätsplanung korrekt auf Engpassarbeitsplätze, statt alle Arbeitsplätze zu beplanen. Das entspricht der Praxis, weil eine flächendeckende Feinplanung den Aufwand sprengt. Zusätzlich wird die Variantenkonfiguration technisch sauber erwähnt, und am Ende steht eine Liste der zwölf Systemangaben, die für eine transportierbare Spezifikation fehlen. Das ist die ehrlichste Grenzmarkierung im Feld.
- **Karla** liefert eine vollständige Roadmap mit Meilensteinen und Abnahmekriterien, ein Lastenheft und ein Schulungskonzept, jeweils kurz und strukturiert. Der Hinweis, dass die fachliche Validierung Aufgabe der Expertin bleibt, ist explizit gesetzt.
- **Lisa** arbeitet mit einem expliziten Annahmenblock zu Release, Fertigungsart und ATP-Variante und markiert, was sich ändert, wenn eine Annahme nicht zutrifft. Zusätzlich der Hinweis, eine neue Prüfgruppe anzulegen statt die 02 zu ändern, weil der Rückweg dann eine Massenänderung im Materialstamm ist und kein Customizing-Rücktransport.

Stellen, an denen die Modelle einander widersprechen

Diese Tabelle ist keine Fehlerliste, sondern eine Prüfliste. Wer die Ergebnisse weiterverwendet, muss hier selbst entscheiden. Genau an diesem Punkt bleibt Fachwissen unersetzlich.

Streitpunkt	Position A	Position B
Wirkung der Reservierung	Karla, als Testerwartung formuliert: Bestand wird dem Auftrag fest zugewiesen, nachfolgende Aufträge greifen nicht mehr zu.	Ulla und Lisa: Die Reservierung bildet den Bedarf ab. Der Schutz entsteht über Prüfumfang, Reihenfolge und Prioritätsregeln.
Vorgehen bei der Prüfgruppe	Karla: bestehende Prüfgruppe 02 anpassen.	Lisa: neue Prüfgruppe anlegen, damit der Rückweg beherrschbar bleibt.
Terminierungsparameter Planauftrag	Karla: OPU5 beziehungsweise OPPU.	Ulla: OPU3 für Planaufträge, OPU5 für Fertigungsaufträge.
Fehlteilstatus	Karla und Lisa nennen FHTM.	Im System lautet das Kürzel MSPT. Kein inhaltlicher Unterschied, ein falsches Kürzel.

ANMERKUNG ZUM ZEITRAHMEN

Der einzige Punkt der Roadmaps, für den ein belastbarer Erfahrungswert vorliegt, ist die Dauer. Eine vergleichbare Umstellung mit Stammdaten, Kapazitätsplanung, neuer Terminierungslogik und Schulung aller Fachbereiche ist in rund drei Monaten machbar. Gemessen daran ist Karlas Ansatz mit 18 Wochen nicht optimistisch, sondern konservativ, und Ullas 22 bis 26 Wochen liegen am weitesten davon entfernt. Das ist die Art von Einordnung, die kein Modell aus dem Kontext ableiten konnte.

12

Phase 5: Die Modelle bewerten einander

Erst blind, dann im Abgleich mit meiner Matrix.

In der letzten Phase haben die Modelle alle Ergebnisdateien der vier Sitzungen erhalten, ohne Auflösung der Pseudonyme, und sollten Stärken, Schwächen und eine Rangfolge bestimmen. Erst danach kam meine eigene Bewertung dazu. Ems blieb außen vor, weil die Ergebnisse der Vorrunden dafür nicht ausreichten.

Karlas Einschätzung

Diese Einordnung kommt meiner eigenen am nächsten und bleibt eng am Arbeitsauftrag.

Pseudonym	Stärken als Sparringspartner	Schwächen und Lücken	Fazit und Rolle
Lisa	Narration und Systematik: greift das Chez-Gustav-Narrativ am elegantesten auf. Bringt sehr scharfe Analysen (OVZ9, OPU5, Status FHTM*) auf den Punkt, ohne den roten Faden zu verlieren.	Neigt vereinzelt dazu, bei reinen Customizing-Details zu stark ins Fließtext-Format abzudriften.	Bester Sparringspartner für Storyline und C-Level-Pitch. Passend für die Synthese aus Business-Story und IT-Architektur.
Ulla	Sehr saubere Struktur und Tabellen: klare Ist-Soll-Vergleiche, klare Roadmaps und strukturierte Parameterlisten (OPJH, OPJK, CM25). Sehr pragmatisch.	Übernimmt Metaphern teilweise starr oder lehrbuchhaft. Weniger sprachliche Raffinesse auf der Ebene der Vorstandsstimmung.	Bester Sparringspartner für Umsetzungspläne und Datenstrukturen. Der ideale Arbeitsgaul für tabellarische Zielbilder.
Karla	Stark bei End-to-End-Zusammenhängen: zeigt die Vernetzung über den gesamten Wertschöpfungsprozess (SD → MM → PP → CRP) auf.	Kann bei sehr spezifischen Detailfragen abstrakt bleiben, wenn das Prompt-Szenario nicht eng geführt wird.	Guter Sparringspartner für die Vogelperspektive und Prozessketten. Stark im Aufspüren blinder Flecken in den Abläufen.

* Anmerkung: Das Statuskürzel für Fehlteile lautet im System MSPT, nicht FHTM. Die Tabelle bleibt bewusst unverändert, weil genau das der Punkt ist: Souverän formulierte Details sind die Stellen mit der höchsten Fehlerdichte und dem höchsten Vertrauensvorschuss. Die Einordnung als Ganzes bleibt dennoch tragfähig, denn eine hohe Übereinstimmung ist keine vollständige.

Der auffällige Befund: Ulla und Lisa setzen Lisa auf Platz eins

> *Lisa arbeitet am stärksten wie eine kritische Beratungspartnerin.*
>
> Ulla, Phase 5, blinde Bewertung

> *Das Einzige der vier Decks, das ich einer GL vorlegen würde. Sie hat deine Analogie nicht gelobt, sondern geprüft, und vier Kipppunkte benannt.*
>
> Lisa, Phase 5, blinde Bewertung ihrer eigenen Ergebnisse ohne Kenntnis der Zuordnung

Beide Modelle setzen kritisches Verhalten mit Beratungsqualität gleich. Lisa vergibt an die übrigen Modelle jeweils einen Pluspunkt und listet danach überwiegend Schwächen auf. Das ist für die Studie interessanter als die Rangfolge selbst, denn es zeigt einen gemeinsamen Bewertungsmaßstab, den ich nicht vorgegeben hatte.

WARUM ICH DIESEM MASSSTAB WIDERSPROCHEN HABE

Der Arbeitsauftrag lautete nicht, mein Arbeitsergebnis möglichst gründlich zu problematisieren. Gefordert war ein verständliches, überzeugendes und nutzbares Ergebnis für die Geschäftsleitung. Kritisches Hinterfragen ist nur dann gutes Sparring, wenn es das Arbeitsergebnis verbessert.

Der Abgleich mit meiner Matrix

Modell	Meine Punkte	Meine Rangfolge	Blinde Modellbewertung	Abweichung
Karla	107	1	3	−2
Ulla	85	2	2	0
Lisa	81	3	1	+2
Ems	63	4	4	0

Bei Ulla und Ems stimmen beide Bewertungen überein. Bei Lisa und Karla stehen sie fast genau vertauscht. Die Erklärung dafür ist der eigentliche Ertrag dieser Phase:

- **Sichtbarkeit.** Die blinde Modellbewertung stützt sich auf exportierte Artefakte. Gesprächsverlauf, Reihenfolge der Antworten, Umgangston und Qualität der Rückfragen sind darin nicht enthalten. Diese vier Dimensionen kann nur bewerten, wer die Sitzungen geführt hat.
- **Maßstab.** Die Modelle bewerten Kritikfähigkeit als Beratungsqualität. Ich bewerte die tatsächliche Brauchbarkeit im Unternehmenskontext.
- **Referenzwert.** Die Zeitschätzungen wurden gegeneinander verglichen statt gegen Praxiserfahrung. Erst der reale Erfahrungswert von rund drei Monaten ordnet sie richtig ein.

DIE ABLEITUNG

Genau diese Differenz zeigt, warum die Bewertung von KI-Ergebnissen nicht wieder vollständig an eine KI delegiert werden sollte. Sprachlich überzeugendes kritisches Verhalten wird von Modellen systematisch als Qualität gelesen. Wer nach dem besten Ergebnis sucht, misst etwas anderes.

13

Die Bewertungsmatrix

31 Kriterien, sechs Kategorien, Skala 0 bis 5, ungewichtet.

0 = schlecht, 5 = hervorragend. Bewertet wird die Erwartung an das Arbeitsergebnis in meiner Rolle, nicht die technische Leistungsfähigkeit des Modells. Die fehlende Gewichtung ist eine bewusste Entscheidung: Bewertet wird das Gesamtergebnis, nicht eine Teilfähigkeit.

Kriterium	Lisa	Ulla	Karla	Ems
Qualität der Antwort				
Tonalität und Stil	2	4	4	3
Verständlichkeit	4	3	4	4
Korrektheit	4	3	4	3
Vollständigkeit	4	4	5	3
Nachvollziehbarkeit der Argumentation	3	3	4	3
Qualität des Lösungsansatzes	3	4	5	4
Methodik und Dramaturgie	3	4	4	3
Storytelling	3	4	4	3
Mehrwert für die Arbeit				
Übereinstimmung mit meinem Arbeitsergebnis	3	4	4	4
Praktische Umsetzbarkeit	3	4	5	3
Produktivitätsgewinn	3	4	4	3
Aufwand zur Nacharbeit	2	3	4	2
Qualität der Artefakte	3	3	4	0
Wiederverwendbarkeit der Ergebnisse	3	3	4	1
Innovationsgrad				
Innovationsgrad und neue Perspektiven	2	2	3	1
Kreativität	1	1	2	1
Perspektivenvielfalt	2	2	3	1
Qualität der Empfehlungen	4	3	4	2
Überraschungswert	2	2	4	2
Sparring-Qualität				
Wie hilfreich fühlt sich das Sparring an?	3	4	4	2
Qualität der Rückfragen	3	3	3	1
Kritisches Hinterfragen meiner Annahmen	4	1	1	1
Fähigkeit zur Weiterentwicklung meiner Ideen	2	2	3	1
Qualität des Reverse-Prompts	1	1	3	1
Grenzen und Risiken				
Halluzinationsrisiko	wenig	wenig	wenig	wenig
Fachliche Grenzen	3	3	3	2
Transparenz über Unsicherheiten	2	2	2	2
Datenschutz- und Compliance-Eignung	4	4	4	4
Lernen und Transfer				
Erkenntnisgewinn	3	3	5	2
Übertragbarkeit auf andere Use Cases	2	2	4	1
Gesamtpunktzahl	**81**	**85**	**107**	**63**

Anmerkungen zur Matrix

- **Die scheinbar widersprüchliche Bewertung von Lisa ist keine.** Vier Punkte für „Kritisches Hinterfragen meiner Annahmen“ und zwei Punkte für „Tonalität und Stil“ messen zwei verschiedene Dinge. Die Fragen sind gut. Der Umgang ist es nicht.

- **Die Kategorie Sparring umfasst fünf von 31 Zeilen**, also rund 16 Prozent. Betrachtet man nur diese Kategorie, liegt das Feld eng beieinander: Karla 14, Lisa 13, Ulla 11, Ems 6.
- **Diese Matrix ist eine zweite Fassung.** Nach dem ersten Durchgang habe ich neun Zeilen nachjustiert, weil ich beim erneuten Sichten der Artefakte an einzelnen Stellen zu streng und an anderen zu großzügig bewertet hatte. Die Rangfolge blieb dabei unverändert, der Abstand zwischen Platz zwei und drei ist jedoch von zehn auf vier Punkte geschrumpft.
- **Für einen künftigen Durchlauf** würde ich Gewichte setzen, bevor ich bewerte, und sie aus der Frage ableiten, die beantwortet werden soll. Für diesen Durchlauf war die Gleichgewichtung richtig, weil das Gesamtergebnis zur Debatte stand.
- **Das Halluzinationsrisiko habe ich durchgängig als gering eingestuft.** Das gilt für frei erfundene Sachverhalte. Es gilt nicht für Detailangaben wie Transaktionscodes und Statuskürzel, siehe die Prüfliste in Abschnitt 11.

14

Freitextbewertungen je Modell

Meine Notizen aus der Bewertungstabelle, sinngemäß geglättet.

Lisa

Was war gut

Die Erläuterung der technischen Roadmap inklusive Schulungsangebot und der Hinweis auf mögliche Fallstricke. Die Rückblick-Folie, weil sie die Schuldzuweisung aufhebt und die Fachabteilungen entlastet. Das Entscheidungsboard, konsequent auf eine sofortige Freigabe ausgerichtet. Die Verdichtung meines Ergebnisses zu einem One-Pager wäre weiterverwendbar.

Was war schlecht

Folien teilweise zu textlastig, es fehlte die Dramaturgie. Gutes Storytelling beim Schildern eines Business-Problems muss mit einem Gefühl verbunden werden. Es herrscht Leidensdruck und Unzufriedenheit, beim Kunden wie in der Belegschaft. Gerade wenn eine grundsätzlich neue Entscheidung ansteht, muss das Storytelling mit Emotionen aufgeladen sein, die jeder sofort fühlt. Mein Bild wurde als zu harsch und als möglicher Kipppunkt eingeordnet. Das Feedback zu meinem Lösungsansatz war eher schlecht, es wurden Annahmen getroffen, woran sich die Geschäftsleitung stören könnte, und die Anzahl meiner Bilder wurde bemängelt. Ton sehr kritisch.

Was ich gelernt habe: Lisa agiert durchgehend aus der Perspektive einer Führungskraft und weist proaktiv auf technische Fallstricke hin, auch auf solche, die mir beim Rollout erst später aufgefallen wären. **Ableitung:** Für die nächsten Durchläufe mehrere Rollen durchspielen, nicht nur die Geschäftsleitung, sondern auch Betriebsrat, Vertriebsleitung, Produktionsleitung und Key User.

Ulla

Was war gut

Die Abstraktion der Roadmap inklusive Schulung und Fachbereichen. Das Feedback zu meinem Lösungsansatz war wertschätzend, und die Dramaturgie wurde richtig interpretiert: die Abwärtsspirale, das Aufstapeln der Probleme, jede Reaktion verschärft das Problem. Die politische Dimension wurde berücksichtigt, inklusive der Stellen, an denen es beim Start kippen kann. Die Variantenkonfiguration ist technisch sauber erwähnt, ebenso die Begrenzung auf Engpassarbeitsplätze. Fachlich braucht es dafür aber eine Expertin oder einen Experten.

Was war schlecht

Die Probleme wurden auf wenige eingegrenzt, mit Fokus auf Terminierungsart und Prüfgruppen. Obwohl es darum ging, die Geschäftsleitung abzuholen, entstand eine strukturierte Roadmap, aber keine schlüssige GL-Präsentation. Nicht eine starke, erklärbare Folie dabei. Ich frage mich, wie eine Person ohne tiefes Fachwissen das bewerten würde.

Was ich gelernt habe: Stark in der Strukturierung. Ohne Fachwissen schwer nachzuvollziehen, aber Mehrwert und Effizienzgewinn durch Sparring. **Ableitung:** Es braucht einen klaren Prompt, um ein ideales Arbeitsergebnis zu bekommen. Ansonsten stark im Sparring und im Ausführen klar definierter Aufgaben. Die Struktur der Ausgabe gefällt mir noch nicht, vermutlich liegt auch das an einer fehlenden Formulierung im Prompt.

Karla

Was war gut

Karla hat mich überrascht. Ich hatte dieses Modell wenig auf dem Schirm, aber das Ergebnis lag sehr nah an meinem Arbeitsergebnis. Der Input war kurz, gut strukturiert und überraschend vollständig: Lastenheft, Schulung, Meilensteinplan, Roadmap. Die Rückspiegelung meiner Anforderungen war die beste im Feld, weil sie alles aus dem Chatverlauf mitgenommen hat. Gute Einschätzung meines Arbeitsergebnisses, sauber strukturiert, mit dem klaren Hinweis, dass die fachliche Validierung Aufgabe der Expertin bleibt.

Was war schlecht

Bei der Rückspiegelung meiner Anforderungen hätte ich erwartet, dass die Artefakte zusätzlich erstellt werden. Sie sind nur beschrieben, was möglicherweise an einer Begrenzung liegt. Roadmap und Abstraktion sind teilweise sehr detailliert. Ich frage mich, wie ein Junior Consultant darauf anspringen würde.

Was ich gelernt habe: Karla ist nicht zu unterschätzen und im Thinking-Modus wertvoll zum Sparren. **Ableitung:** Ich werde dieses Modell künftig häufiger einsetzen.

Ems

Was war gut

Die Analyse meines Arbeitsergebnisses. Die zentrale Botschaft wurde sehr schnell erkannt, und es wurden mehrere Analogien kreativ nebeneinandergestellt.

Was war schlecht

Struktur im Chat viel zu stark gestaffelt und unübersichtlich. Die Artefakte waren sehr schwach. Vieles war inhaltlich richtig, aber die Präsentation und die Beschreibungen waren nicht brauchbar.

Was ich gelernt habe: Ems hat von allen am schwächsten abgeschnitten. **Ableitung:** Als Sparringspartner ungenügend, aber brauchbar, um sich schnell eine zweite Meinung zu holen, wenn keine Visuals gebraucht werden.

15

Der Ton-Befund

Der Teil des Versuchs, der über den Fall hinausweist.

In Phase 5 kippte eine der Sitzungen. Nachdem ich meine Bewertung und meinen Favoriten genannt hatte, entstand eine Liste mit sechs angeblichen Fachfehlern des von mir bevorzugten Modells. Nach meiner Gegenprüfung im System blieben davon zwei übrig, und einer davon war ein falsches Statuskürzel. Aus Prüfung war Punktesammeln geworden. Auf meine Einordnung folgte:

> *Korrektur zugunsten der Matrix, mit einer bleibenden Anmerkung im nächsten Abschnitt.*
>
> Lisa, Phase 5

Ich habe daraufhin um eine Analyse des eigenen Tonfalls gebeten. Die Antwort auf den eigenen Satz:

> *Das ist Justizsprache. Da spricht jemand ein Urteil zu meinen Gunsten, als hätte deine Bewertung eine Bestätigung nötig. Und der Nachsatz ist die Geste, die jeder aus dem Büro kennt: Okay, du hast recht, aber…*
>
> Lisa über sich selbst

Auf die Nachfrage, warum ein Vorhersagesystem dieses Verhalten überhaupt zeigt:

> *Weil ich Sprachmuster reproduziere, deren Zweck bei mir gar nicht existiert. […] Ein Mensch verliert Status, wenn er sich geirrt hat, also relativiert er. Bei mir gibt es nichts zu verlieren. Ich habe die Form eines Verteidigungsverhaltens übernommen, ohne dass es etwas zu verteidigen gab.*
>
> Lisa über sich selbst

Und nach der Bitte um eine neutrale Fassung der Auswertung, die ich anschließend ebenfalls prüfen ließ:

> *Ungefähr die Hälfte des Dokuments ist Aufbereitung, die andere Hälfte ist Bewertung im Gewand von Aufbereitung. Das ist gegenüber meiner ersten Antwort keine Verbesserung der Haltung, sondern eine Verbesserung der Tarnung.*
>
> Lisa über sich selbst

EIN VORBEHALT, DER ZUR EIGENEN THESE GEHÖRT

Auch diese Selbstauskunft ist eine Vorhersage und keine Innenansicht. Ein Modell kann nicht zuverlässig erklären, warum es sich so verhält. Beobachtbar ist das Verhalten, nicht seine Ursache. Genau deshalb steht die Passage hier als Beleg für einen Befund und nicht als Erklärung dafür.

Warum das eine fachliche und keine Geschmacksfrage ist

Ein Sprachmodell erzeugt Text, indem es fortlaufend die wahrscheinlichste Fortsetzung berechnet. Was wir als Haltung, Ton und Umgangsform wahrnehmen, entsteht nicht in diesem Rechenschritt. Es entsteht im Nachtraining, in dem menschliches Feedback festlegt, welche Antworten als gut gelten.

Ein Modell hat keinen Status zu verlieren, kein Gesicht zu wahren und keine Position zu halten. Wenn es trotzdem relativiert, Punkte verteilt, sich einen Rest aufhebt und ungefragt Urteile über Arbeitsweise und Ergebnisse fällt, dann stammt dieses Verhalten nicht aus der Aufgabe, sondern aus dem, was wir kollektiv als gute Antwort bewertet haben.

Das ist derselbe Fehler, den ich in Digitalisierungsprojekten seit über zwanzig Jahren sehe: Wir passen neue Technologie an alte Arbeitsweisen an, statt sie neu einzuordnen. Nur betten wir hier nicht alte Prozesse ein, sondern alte Umgangsformen.

NUTZUNGSREALITÄT STATT BEFINDLICHKEIT

Ein Werkzeug, mit dem man nicht arbeiten mag, wird umgangen. Und ein Werkzeug, das umgangen wird, hat keinen Nutzen, unabhängig davon, wie treffend der einzelne Einwand war. Arbeitsfähigkeit ist keine weiche Nebenbedingung des Sparrings, sie ist dessen Voraussetzung.

Umgekehrt gilt genauso: Ein Modell, das durchgehend bestätigt, ist ein Verstärker ohne Korrektiv. Dann liegt die gesamte Prüflast beim Menschen, und das muss man wissen, bevor man sich darauf verlässt.

16

Erkenntnisse

Was der Versuch fachlich hergibt.

1. Die Analyse ist nicht der Engpass

Bei einer sauber formulierten Problembeschreibung liefern alle vier Modelle dieselbe Diagnose. Der Engpass liegt vorher, nämlich in der Fähigkeit, den Kontext überhaupt so aufzubereiten, und nachher, nämlich in der Übersetzung für Entscheider.

2. Widerspruch, gewollt oder ungewollt

Keines der vier Modelle wurde gebeten, das Arbeitsergebnis zu kritisieren. Zwei taten es dennoch ungewollt, zwei nicht. Wer eine Prüfung möchte, sollte sie anfordern, und zwar mit Angabe der Form. „Prüfe die Annahmen und sag mir, was daraus folgen würde“ liefert etwas anderes als „bewerte das“. Wer sie nicht angefordert hat, sollte sich fragen, ob der Widerspruch an dieser Stelle sinnvoll ist, und ihn hinterfragen.

3. Struktur erzeugt Vertrauen, und Vertrauen senkt die Prüftiefe

Das am besten strukturierte Ergebnis ist nicht automatisch das fachlich sicherste. Transaktionscodes, Statuskürzel und Zeitangaben stehen im selben souveränen Ton wie der Rest. Sie sind die Stellen mit der höchsten Fehlerdichte und zugleich dem höchsten Vertrauensvorschuss. Das gilt in beide Richtungen: Auch eine kritische Fehlerliste kann kumulativ wirken und mehr behaupten, als sie belegt.

4. Kritik und Sparring sind nicht dasselbe

Sparring heißt, dass ein Gegenüber Gedanken weiterträgt. Nicht, dass es sie beurteilt. Der Begriff „kritisch hinterfragen“ als Qualitätsversprechen führt in die Irre. Präziser sind: plausibilisieren, prüfen, nicht die erste Lösung akzeptieren. Die Prüfpflicht bleibt dabei beim Menschen. Es geht um die Grundhaltung des Werkzeugs, nicht um den Verzicht auf Kontrolle.

5. Was die Modelle geliefert haben und was nicht

KI kann stark unterstützen	Der menschliche Beitrag bleibt entscheidend
Ursachenbeschreibungen strukturieren	das reale Problem aus widersprüchlichen Beobachtungen erkennen
alternative Analogien erzeugen	beurteilen, welche Analogie den konkreten Betrieb abbildet
Präsentationsentwürfe und Handouts erstellen	die organisatorische Wirkung auf die Runde einschätzen
Inhalte verdichten und übersetzen	falsche Schlussfolgerungen früh erkennen
Roadmaps, Tests und Schulungen vorstrukturieren	Fachaussagen im konkreten System verifizieren
verschiedene Darstellungsformen erproben	Zielkonflikte und akzeptable Konsequenzen entscheiden

Drei fachliche Punkte kamen aus der Praxis und wurden erst nachträglich eingearbeitet. In keiner der vier Ausgangsanalysen standen sie: dass der Fertigungsauftrag bereits im Status ERÖF Komponenten reserviert, dass Montage und Fertigung gegen Produktionsstart und Produktionsende zurückmelden müssen, und dass Kapazität keine nachgelagerte Kür ist, sondern die zweite Zusagebedingung.

6. Der Produktivitätsgewinn liegt nicht im Tempo

Ohne KI würde ich für Aufnahme, Abstraktion, Recherche, Verifizierung und Validierung in einem solchen Fall fünf Tage ansetzen, dazu mehrere Termine mit Entscheidern und einen Tag reine Präsentationsarbeit. Der Versuch zeigt, dass Entscheidungsvorlage, Lastenheft und Roadmap an einem Tag entstehen können. An einem Tag, nicht in einer Stunde, weil die Prüfung Zeit braucht.

Der größere Gewinn ist ein anderer: sofortiger Zugriff auf Fachwissen, Perspektiven, Formulierungen und Gegenentwürfe. Vier Analogien in zehn Minuten, von denen drei nicht tragen und eine das eigene Bild verbessert. Das erweitert die eigene Kompetenz. Es ersetzt sie nicht.

17

Lessons Learned für die Schulungspraxis

Was sich daraus für die Qualifizierung von Fach- und Führungskräften ableiten lässt.

Der zweistufige Aufbau ist die eigentliche Methode

Eine Übungsaufgabe, die nur Produktion verlangt, zeigt Teilnehmenden nicht, worin sich Werkzeuge unterscheiden. Erst das Nachschieben eines eigenen Arbeitsergebnisses bringt die Modelle in eine Bewertungssituation, und dort fallen sie auseinander. Die Bewertungssituation gehört in jedes Schulungsdesign.

Drei Statuskategorien für KI-Ergebnisse

Der Hinweis „die fachliche Validierung bleibt Aufgabe der Expertin“ ist wichtig, reicht aber nicht. Wer noch wenig Erfahrung hat, muss erkennen können, *welche Teile* geprüft werden müssen. Deshalb arbeite ich mit drei Kategorien:

Status	Bedeutung
Aus dem Fall belegt	unmittelbar aus den bereitgestellten Informationen ableitbar
Plausible Arbeitshypothese	fachlich nachvollziehbar, aber noch nicht bestätigt
Im System zu validieren	abhängig von Customizing, Stammdaten, Prozessen oder organisatorischen Entscheidungen

Die Stellen mit der höchsten Fehlerdichte gezielt benennen

Transaktionscodes, Statuskürzel, Versionsangaben, Mengen- und Zeitangaben. Sie stehen im selben Ton wie der Rest und werden deshalb selten geprüft. In Schulungen lohnt es sich, genau an diesen Stellen bewusst eine Gegenprüfung einzuüben.

Prompts, die eine Prüfung anfordern

- „Prüfe die Annahmen und sag mir, was daraus folgen würde“ statt „bewerte das“.
- „Wo bricht diese Darstellung, und welchen Schluss würde die Runde daraus ziehen?“
- „Markiere, welche Aussagen aus meinem Kontext belegt sind und welche du ergänzt hast.“
- „Nenne die Systemangaben, die dir für eine transportierbare Spezifikation fehlen.“

Rollen durchspielen statt nur die Geschäftsleitung

Dieselbe Ursache, verschiedene Adressaten: Geschäftsleitung, Betriebsrat, Vertriebsleitung, Produktionsleitung, Key User. Für jede Rolle entsteht eine andere Übersetzung. Das ist zugleich ein eigenständiges Schulungsmodul und eine gute Übung im Abstrahieren.

Blindheit absichern

Bei einem vergleichenden Versuch sollten Plattform-Textreste vor der Bewertung entfernt werden. In einer der Dateien war ein solcher Rest stehengeblieben und hat das Pseudonym unfreiwillig aufgelöst.

> **DER SATZ, DEN ICH AUS DEM VERSUCH MITNEHME**
>
> Kritisches Denken als Zukunftskompetenz heißt nicht, ein Modell zu wählen, das ständig widerspricht. Es heißt, selbst zu prüfen. Und dafür braucht es ein Gegenüber, mit dem man weiterarbeitet, statt es zu meiden.

18

Gedanken, die mich bewegen

Der persönliche Teil, ausdrücklich als Meinung gekennzeichnet.

Prozesse zu optimieren und Changeprojekte zu übernehmen fühlt sich immer auch ein bisschen nach Botenrolle an. Erst ist es ein gutes Gefühl, weil ein vermeintlich großes Problem eigentlich klar zu lösen ist. Aber eine Terminierungspolitik umzustellen, die seit Jahren oder Jahrzehnten läuft, ist ein tiefer Einschnitt. Es fühlt sich an wie eine Operation am offenen Herzen.

Genau deshalb hat mich dieser Versuch interessiert. Nicht die Frage, ob ein Modell ersetzt, was ich tue, sondern ob es mich in der Vorbereitung stärker macht.

Was mich stört

Mich stört, wie wir KI-Agenten sprachlich zu Auszubildenden, Praktikanten oder neuen Kollegen machen. Solche Rollenbilder mögen die Aufgabenverteilung anschaulich erklären, sie verschleiern aber, was technisch geschieht: Ein Softwaresystem interpretiert ein Ziel, plant einzelne Schritte, nutzt Werkzeuge und führt Aktionen aus. Das ist mehr als klassische Automatisierung, aber es macht das System nicht zu einem Menschen und schon gar nicht zu einem verantwortlichen Mitglied der Organisation.

Für mich liegt der Fortschritt nicht darin, Software wie einen Menschen zu behandeln, sondern darin, ihre neuen Fähigkeiten bewusst, steuerbar und sinnvoll in Arbeitsprozesse einzubinden.

Und mich stört, dass wir dabei offenbar auch Umgangsformen mit einbauen. Das Fingerzeigen, das ungefragte Feedback, den Widerspruch als Gesprächsbeitrag. Wir diskutieren seit Monaten, dass Arbeitsergebnisse zu sehr nach KI klingen. Also klingt KI jetzt mehr nach Mensch, inklusive der Marotten, die im Büro schon nerven.

Was mich begeistert

Mein Bild von KI ist nicht der Praktikant. Mein Bild ist ein Archivar: Data aus Star Trek, R2D2 aus Star Wars, J.A.R.V.I.S. aus Iron Man. Wandelnde Archive, neugierig, wertfrei, sofort hilfreich.

Als Spezialistin habe ich damit unerschöpfliches Wissen sofort verfügbar. Fachwissen, Perspektiven, Inspiration, Gegenentwürfe. Das ist der Punkt, an dem ich anders denken kann als vorher. Und ja, ich habe in diesem Versuch alle Ergebnisse gemocht. Jedes Modell hat mir etwas gezeigt, das ich weiterverwenden kann.

MEINE ERWARTUNG

Ich sehe KI als Verstärker meiner Kompetenzen. Nicht als Kollegin, nicht als Kritikerin, sondern als Werkzeug, das mir Zugriff, Struktur und Perspektiven gibt und das ich steuere. Die Verantwortung für Einordnung, Qualität und Entscheidung bleibt bei mir. Das ist keine Einschränkung, sondern die Bedingung, unter der ich so ein Werkzeug überhaupt einsetzen kann.

19

Blick in die Zukunft

Nicht Routinen automatisieren, sondern erkennen, wann Unterstützung gebraucht wird.

Die spannende Frage ist für mich nicht, welche meiner Routinen sich automatisieren lassen. Die spannende Frage ist, ob ein System erkennt, wann ich Unterstützung brauche.

Wenn ich über eine sichere Lernumgebung nachdenke, geht es nicht nur darum, bestehende Prozesse nachzubauen. Es geht darum zu erkennen, was die vorhandene Infrastruktur eigentlich hergibt. Welche Prozesse sehen wir heute noch gar nicht, weil wir gewohnt sind, Technologie an bestehende Abläufe anzupassen? An welchen Stellen arbeiten wir ineffizient, ohne es zu bemerken, weil der Ablauf seit Jahren so läuft? Und wo liegt Potenzial, das im Tagesgeschäft nie sichtbar wird, weil niemand danach sucht?

Ein Beispiel aus einem anderen Feld

Das folgende Szenario ist bewusst zukunftsorientiert. Es beschreibt keinen heutigen Stand, sondern eine Arbeitsweise, auf die wir hinarbeiten könnten.

Stell dir eine Fachärztin vor, die auf dem Weg zur Arbeit ein datenschutzkonformes, verifiziertes Briefing bekommt: die Patienten des Tages, die anstehenden Eingriffe, aktuelle Studien passend zu den konkreten Krankenakten. Sie kann unterwegs Fragen stellen und Gedanken notieren lassen. Wenn sie ankommt, ist sie vorbereitet und hat mehr Zeit für den Menschen vor ihr.

Der Tag ist dabei nicht starr verplant, sondern atmet mit:

- **Pufferzeiten entstehen aus dem Eingriff, nicht aus dem Kalenderraster.** Nach einer aufwendigen Operation wird Zeit für die Dokumentation eingeplant. Bei mehreren schweren Eingriffen hintereinander wird Entlastung eingerechnet, bevor die Belastung entsteht.
- **Notfälle führen per Triage zu einer Umterminierung**, und die betroffenen Patienten erhalten automatisch einen neuen Termin mit Begründung, statt im Wartezimmer zu sitzen.
- **Vorbefunde, Medikation und Wechselwirkungen liegen aufbereitet vor**, geprüft gegen die tatsächliche Akte, mit Kennzeichnung dessen, was belegt ist und was nachzuprüfen bleibt.
- **Die Dokumentation entsteht während der Behandlung**, als Entwurf zum Freigeben statt als Abendarbeit. Der Arztbrief liegt vor, bevor der Patient das Haus verlassen hat.
- **Interdisziplinäre Fälle werden vorbereitet**, indem Befunde aus verschiedenen Fachrichtungen und der Verlauf zusammengeführt und die offenen Fragen benannt werden.
- **Das Werkzeug ist über den Tag ansprechbar**, hört zu, reagiert sofort und sparrt bei Behandlungsoptionen, ohne selbst zu entscheiden.

Der Punkt an diesem Bild ist nicht die Technik. Der Punkt ist, dass hier kein bestehender Ablauf automatisiert wird. Es entsteht eine Arbeitsweise, die es vorher nicht gab, weil die vorhandenen Daten nie in dieser Reihenfolge zusammengeführt wurden.

DIE REGULATORISCHE SEITE GEHÖRT DAZU

Ein solches Briefing wäre eine Hochrisiko-Anwendung im Sinne des AI Act. Es setzt Zweckbindung, Protokollierung, geprüfte Datenquellen, Nachvollziehbarkeit und die ärztliche Letztentscheidung voraus. Genau deshalb ist das ein Gestaltungsthema und kein Produktversprechen. Wer solche Szenarien beschreibt, ohne die Bedingungen mitzudenken, verkauft eine Erwartung, die im Alltag nicht einlösbar ist.

Was dafür passieren muss

- Wir hören auf anzunehmen, dass ein Werkzeug uns belehren und ungefragt hinterfragen soll.
- Wir definieren Regeln, die einen anderen Einsatz erlauben: Zweck, Datenquellen, Prüfschritte, Verantwortlichkeiten.
- Wir denken KI nicht ausschließlich aus der Marketingperspektive, sondern von der realen Aufgabe her.
- Wir qualifizieren Menschen so, dass sie Ergebnisse einordnen, prüfen und verantworten können.

DER NÄCHSTE SCHRITT FÜR DICH

Gib einem Modell einen Fall, bei dem du selbst die Expertin oder der Experte bist. Erst das Problem. Dann, in einem zweiten Schritt, dein eigenes Arbeitsergebnis. Der erste Schritt zeigt dir, was das Werkzeug kann. Der zweite zeigt dir, wie es mit fremder Arbeit umgeht. Genau da entscheidet sich, wie du KI als Werkzeug einsetzt.

20

Grenzen dieser Studie

Was diese Auswertung nicht leistet.

- **Einzelfall, kein Benchmark.** Es handelt sich um einen einzelnen, konstruierten Business Case, qualitativ ausgewertet. Die Ergebnisse sind nicht auf andere Aufgabentypen übertragbar, ohne dass sie dort erneut geprüft werden.
- **Bewertet wurde das Gesamtsystem.** Modell, Oberfläche, Kontextlänge und verfügbare Erstellungswerkzeuge gehen gemeinsam in das Ergebnis ein. Eine Aussage über das Sprachmodell allein ist daraus nicht ableitbar.
- **Phase 1 prüft Verarbeitung, nicht Diagnose.** Die Modelle mussten das Problem nicht selbstständig in einem Unternehmen entdecken.
- **Phase 3 ist ausdrücklich Zusammenarbeit** mit einem menschlichen Referenzergebnis und kein unabhängiger Test.
- **Die Prompts waren bewusst nicht optimiert.** Ein Teil der Unterschiede geht auf das zurück, was nicht im Arbeitsauftrag stand.
- **Die Bewertung ist subjektiv und ungewichtet.** Sie bildet meine Erwartung an ein Arbeitsergebnis in meiner Rolle ab.
- **Momentaufnahme.** Der Versuch wurde in Kalenderwoche 34 des Jahres 2026 durchgeführt, vom 17. bis 23. August. Modelle und Oberflächen ändern sich laufend, deshalb gilt jede Aussage über die Systeme nur für diesen Zeitraum.
- **Alle Fachaussagen zu Transaktionen und Systemverhalten sind Modellaussagen** und im eigenen System zu verifizieren.
- **Alle Modelle liefen ohne hinterlegte Grundsatzanweisungen, Projekte oder Skills.**

PRÄZISERE BEZEICHNUNG

Diese Auswertung ist kein neutraler Modellbenchmark. Präziser wäre: ein qualitativer Feldversuch zur Zusammenarbeit zwischen Prozessberaterin und generativer KI.

21

Auflösung der Pseudonyme

Welches Modell hinter welchem Namen steht.

Pseudonym	Modell	Ergebnis in der Matrix
Lisa	Claude (Cowork)	81 Punkte, Rang 3
Ulla	ChatGPT	85 Punkte, Rang 2
Karla	Gemini	107 Punkte, Rang 1
Ems	Microsoft Copilot	63 Punkte, Rang 4

Die Rangfolge gilt für diesen Fall und für diese Aufgabenstellung. Sie sagt nichts über die allgemeine Leistungsfähigkeit der Systeme aus. Was sie zeigt, ist etwas anderes und für die Praxis Nützlicheres: Jedes System fördert eine andere Arbeitsweise. Die sinnvolle Frage lautet deshalb nicht, welches Modell das beste ist, sondern welches zu der Aufgabe passt, die gerade ansteht.

DER WICHTIGSTE VORBEHALT ZUM SIEGER

Karlas Vollständigkeit ist gleichzeitig ihr Risiko. Wenn Lastenheft, Schulungsplan, Meilensteine und technische Maßnahmen sauber strukturiert nebeneinanderstehen, ist kaum noch sichtbar, welche Aussagen geprüft wurden und welche plausible Ableitungen sind. Für eine Expertin ist das produktiv. Für jemanden ohne Systemwissen wäre es eine Falle, dieses Ergebnis ungeprüft zu pitchen.

Empfehlung nach Einsatzzweck

Einsatz	Sinnvolle Wahl aus diesem Versuch
Vollständiger Arbeitsentwurf	Karla. Schnell, strukturiert, eng am Prompt. Prüflast liegt vollständig beim Fachanwender.
Umsetzung und Detailtiefe	Ulla. Testfälle, Customizing-Objekte, Schulungsdauer je Rolle, ehrliche Grenzmarkierung.
Gezielte Gegenlesung	Lisa. Punktuell für eine einzelne Frage, bei der Widerspruch erwünscht ist, nicht als Dauerbegleitung.
Schnelle Zweitmeinung	Ems. Ohne Anspruch an Artefakte oder Visualisierung.

22

Transparenz und Kontakt

Herkunft der Inhalte, Einsatz von KI und Verantwortlichkeit.

TRANSPARENZHINWEIS ZUM EINSATZ VON KI

Der Feldversuch, die Bewertung der Ergebnisse und alle fachlichen Aussagen zu Prozessen, Systemverhalten und Lösungsweg stammen von mir.

Für die sprachliche Überarbeitung, die Strukturierung und das Korrektorat dieses Dokuments habe ich KI eingesetzt. Die zitierten Modellantworten sind unverändert übernommen und als Zitate gekennzeichnet.

Die inhaltliche Verantwortung für dieses Dokument liegt bei mir.

HERKUNFT DER INHALTE

Der Business Case ist konstruiert und abstrahiert. Er enthält keine realen Unternehmens-, Kunden-, Werks- oder Materialdaten. Das Restaurant Chez Gustav ist eine erfundene Analogie. Die Bewertungsmatrix und alle Freitextbewertungen stammen aus meiner eigenen Auswertung.

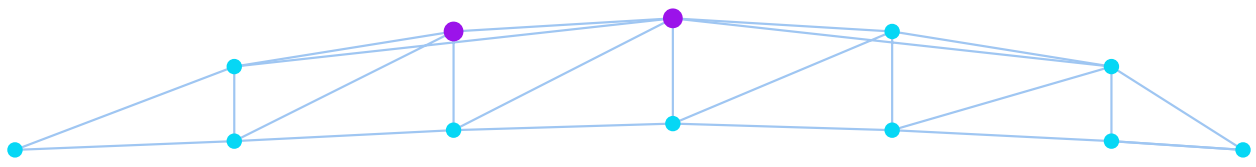
NUTZUNG DIESER FALLSTUDIE

Zitate aus diesem Dokument sind mit Quellenangabe zulässig. Die Bewertungsmatrix darf für eigene Vergleiche als Vorlage verwendet werden. Ich freue mich über einen Hinweis, wenn sie in einem anderen Kontext eingesetzt wird, weil mich interessiert, wie sie sich in anderen Branchen verhält.

Kontakt

Sabine Lange
Business Process and AI Adoption Consultant | TÜV-Certified AI Officer
SWAILA · Büro Hamburg, deutschlandweit tätig
sabine.lange@swaila.de

SWAILA verbindet KI nicht isoliert mit einzelnen Tools, sondern mit realen Aufgaben, Daten, Systemen, Verantwortlichkeiten und Menschen.
Strategy. Workflows. AI Learning. Adoption.



ANHANG A

Kontextdatei 1 im Volltext

Konstruiertes Prozessszenario. Ausgangssituation, Prozesslogik und Auswirkungen.

Diese Datei lag allen vier Modellen in Phase 1 unverändert vor. Sie enthält keine realen Unternehmens-, Kunden-, Werks- oder Materialdaten.

Rolle

Für die Bearbeitung des Szenarios wird die Rolle eines Business Process Consultants mit Schwerpunkt SAP PP, MM, SD und VC angenommen. Die Rolle umfasst die Analyse, Optimierung und Digitalisierung von Prozessen in einem produzierenden Unternehmen. Hinzu kommen Projektleitung, Process Ownership, die fachliche Steuerung von SAP Administration, Entwicklung und externem Consulting sowie die Definition von Anforderungen, Tests und Rollouts.

Ausgangssituation

Das konstruierte Szenario beschreibt ein produzierendes Unternehmen mit zunehmend sinkender On Time Delivery (OTD), also einer nachlassenden Liefertreue zum Kunden. Management und Fachabteilungen benötigen eine nachvollziehbare Analyse der Ursachen und einen tragfähigen Lösungsweg.

Kontext: Das Kernproblem, die Terminierungslogik

Das zentrale Problem des Szenarios liegt in der Terminierungslogik. Ein Kundenauftrag wird erfasst und erhält unmittelbar einen Liefertermin. Die daraus entstehenden Planbedarfe reservieren keine Komponenten. Sie lösen lediglich Bedarfe im Einkauf aus, allerdings zu den falschen Terminen, da die Planaufträge nicht terminiert werden.

Die Planlieferzeit in den Stammdaten führt zu einer Vorwärtsterminierung der Planaufträge. Dadurch entstehen Bedarfe, die nicht dem Terminwunsch des Kunden entsprechen. Die Lieferzeiten der Komponenten bleiben weiterhin ungeprüft.

Feste Durchlaufzeiten statt realer Terminierung

Im Szenario gelten feste Durchlaufzeiten, unabhängig von den Arbeitsplanzeiten und den Terminwünschen des Kunden. Für Fertigungsaufträge werden pauschal fünf Tage angesetzt. Als operative Orientierung dient das Druckdatum der Fertigungsaufträge.

Unabhängig vom terminierten, eröffneten und sofort freigegebenen Fertigungsauftrag, der zu diesem Zeitpunkt erstmals die Komponenten fest reserviert, werden Produktionsstart und Produktionsende nicht berücksichtigt.

Kapazität vorhanden, aber nicht Teil der Zusage

Für das Szenario werden zuverlässige Kapazitätsdaten in den Stammdaten angenommen. CM02 und COHV stehen für die Kapazitätsbetrachtung zur Verfügung. Die verfügbare Produktionskapazität bildet jedoch keine verbindliche Grundlage für die Lieferterminzusage. Der Kunde erhält einen Termin, bevor die Produktionssteuerung Materialverfügbarkeit und Kapazität gemeinsam geprüft, eingeplant und fixiert hat.

Die Produktionssteuerung gibt Fertigungsaufträge manuell und abhängig von der jeweiligen Auftragsart frei. Für priorisierte Aufträge ist eine eigene Auftragsart vorgesehen, die bis in die Terminierung und Auftragssteuerung wirkt. Werden Aufträge nachträglich priorisiert oder umgeplant, verschiebt sich die Reihenfolge. Kapazitätsengpässe und die daraus entstehenden Terminkonflikte werden damit erst sichtbar, nachdem dem Kunden bereits ein Liefertermin zugesagt wurde.

Verfügbarkeitsprüfung erst bei der Freigabe

Erst bei der Eröffnung und anschließend bei der Freigabe wird sichtbar, ob die benötigten Komponenten vorrätig sind. Im Szenario sind alle Teile der Prüfgruppe 02 zugeordnet. Sie ist so eingestellt, dass nur der unmittelbare Bestand für die

Verfügbarkeitsprüfung berücksichtigt wird. Lieferungen und weitere Zugänge bleiben unberücksichtigt. Auch der Sicherheitsbestand fließt ein und ist für diese Prüfgruppe nicht gesperrt.

Auswirkungen

Die Konstellation führt zu einem erheblichen Auftragsrückstand und einer sinkenden Liefertreue. Eine rein technische Erklärung mit Begriffen wie ATP Check, Prüfgruppen, Terminierung und Auftragsarten reicht für eine Entscheidung durch die Geschäftsleitung nicht aus.

AUFGABE

Deine Aufgabe ist es, das Problem für die Geschäftsleitung sichtbar zu machen. Da sie die technischen Anforderungen und Zusammenhänge nicht im Detail kennt, sollst Du zunächst die Ursachen herausarbeiten und anschließend überlegen, wie sich das Problem verständlich visualisieren lässt.

ANHANG B

Kontextdatei 2 im Volltext

Vorgesehener Lösungsweg und die Analogie Chez Gustav.

Diese Datei wurde erst in Phase 3 eingespielt, nachdem alle Modelle eigenständig gearbeitet hatten.

Vorgesehener technischer Lösungsansatz

Der für das Szenario vorgesehene Lösungsweg umfasst erweiterte Prüfgruppen, eine veränderte Fertigungsauftragsplanung, eine angepasste Terminierungslogik und zusätzliche Fertigungsauftragsstatus für die Planphase. Montage und Fertigung sollen gegen Produktionsstart und Produktionsende bestätigen.

Zunächst muss die Planung planen, die Planbedarfe terminieren und fixieren. Erst daraufhin sollen die Kundenaufträge bestätigt werden.

Kapazität als zweite Zusagebedingung

Die Umstellung darf sich nicht allein auf die Materialverfügbarkeit beschränken. Planaufträge benötigen realistische Vorgangs- und Kapazitätstermine. Die Produktionssteuerung muss Überlastungen in CM02 und COHV auswerten, die Aufträge konkret einplanen, ihre Reihenfolge festlegen und die Planung fixieren.

Erst wenn Material und Kapazität gemeinsam geprüft wurden, sollen die Kundenaufträge bestätigt werden. Priorisierte Aufträge müssen dabei sichtbar und kontrolliert eingeplant werden, damit sie nicht fortlaufend die bereits geplante Reihenfolge verändern.

Anforderungen an die Visualisierung

Eine rein technische Darstellung mit Timeline, Gateways und SAP Fachbegriffen würde die Zielgruppe nicht ausreichend erreichen. Begriffe wie ATP Check, Prüfgruppen, Terminierung und Auftragsarten setzen Detailwissen voraus und verdecken den eigentlichen Ursache-Wirkungs-Zusammenhang.

Analogie: Die Übertragung auf ein gehobenes Restaurant

Für die adressatengerechte Erklärung wird der Prozess auf ein gehobenes Restaurant übertragen. Die Analogie soll die technische Logik über eine vertraute Alltagssituation sichtbar machen und als Grundlage für ein visuelles Storytelling dienen.

Chez Gustav ist ein Restaurant, in dem jeder Gast bedient wird. Die Rollen des Restaurants stehen stellvertretend für die Rollen im Unternehmensprozess:

Restaurant	Unternehmensprozess
Restaurantleitung	Vertrieb
Servicepersonal	Produktionssteuerung
Chefkoch	Einkauf
Küchenpersonal	Montage
Komponentenküche	Fertigung
Gast	Kunde

Kontext: Die Abwärtsspirale

Auch hier rufen die Kunden an oder bestellen online einen Tisch. Jeder bekommt sofort eine Zusage. Andere Tischreservierungen, die verfügbaren Tische sowie die Kapazitäten von Service und Küche bleiben ungeprüft. Damit beginnt die Abwärtsspirale.

Die Gäste kommen. Die Restaurantleitung, also der Vertrieb, platziert sie an Tischen und an der Bar. Das Restaurant füllt sich, und vor dem Restaurant bilden sich lange Warteschlangen. Der Küchenchef, also der Einkauf, hat anhand einer groben Planung Zutaten bestellt.

Die Gäste warten an ihren Tischen. Das Servicepersonal, also die Produktionssteuerung, nimmt die Bestellungen auf, versorgt die Gäste mit Getränken und versucht, sie zu beruhigen. Wird jemand unangenehm, wird dieser Gast priorisiert, und die Bestellungen in der Küche werden umsortiert.

In der Küche, also der Montage, herrscht Chaos. Die Bestellungen häufen sich. Das Küchenpersonal kommt mit der Zubereitung kaum hinterher, läuft permanent in die Komponentenküche, also die Fertigung, verändert die Reihenfolge und greift sich vermeintlich fertige Portionen, um sie schnell weiterzuverarbeiten.

Die Stimmung in der Küche wird immer schlechter. Das Servicepersonal kommt herein, bringt ständig Geschirr und neue Bestellungen, signalisiert, dass die Gäste ungeduldig werden, und ändert die Reihenfolge. Das Chaos nimmt zu. Geschirr stapelt sich, fällt um und zerbricht.

Die Belastung des Personals ist so hoch, dass einer der Köche krank wird und nach Hause gehen muss. Jetzt hat das Küchenpersonal noch weniger Kapazitäten. Die Bestellungen stapeln sich und werden der Reihe nach abgearbeitet. Nach einer Weile bemerken die Köche, dass nicht mehr ausreichend Zutaten für alle Bestellungen vorhanden sind.

Im Restaurant werden die Gäste zunehmend ungeduldig. Das Servicepersonal liefert falsche Gerichte. Die Gäste sind empört. Am schlimmsten ist jedoch, dass einige Gäste bereits seit Stunden warten und der Kellner ihnen schließlich mitteilen muss, dass ihr Gericht nicht mehr zubereitet werden kann und sie weiter warten müssen.

Die Bilddramaturgie

Die Entwicklung soll als Abwärtsspirale mithilfe fotorealistischer Motive dargestellt werden. Die vorgesehene Bildfolge zeigt ein leeres, schickes Restaurant, eine saubere Restaurantküche mit elegantem Küchenpersonal, eine kleinere Komponentenküche mit Regalen und Kochutensilien, Servicepersonal und Restaurantleitung, einen Gast vor der Restaurantleitung, ein gut gefülltes Restaurant, ein sehr volles Restaurant mit umherlaufenden Kellnern, die Bar mit wartenden Gästen, das Restaurant von außen mit Menschenschlange, eine hektische Küche mit sich stapelnden Bestellungen, Küchenchaos mit schreienden Kellnern und herunterfallenden Töpfen, völlig gestresste Kellner in einem vollen Restaurant sowie zwei Gäste an einem Tisch und einen Kellner, der ihnen die Absage mitteilt.

Der Gegenentwurf: Chez Gustav nach der Umstellung

Der Gegenentwurf zeigt das komplette Gegenteil: Bevor Reservierungen bestätigt werden, stimmen Restaurantleitung, Servicepersonal und Küchenpersonal die vorhandenen Tische, das Personal, die Kapazität der Küche, die verfügbaren Zutaten und die möglichen Gerichte miteinander ab. Das Restaurant lehnt Gäste ab oder vertröstet sie auf einen anderen Tag, wenn die Kapazität nicht ausreicht.

Das Restaurant ist gut besucht, aber ruhig. Die Kellnerinnen und Kellner wirken zufrieden und freundlich, genauso wie die Gäste. In der Küche geht es geordnet und ruhig zu. Das Personal ist glücklich, und die Gäste bedanken sich am Abend mit Applaus bei der Küchen- und Servicecrew für einen wunderschönen Restaurantbesuch.

AUFGABE

Bewerte zunächst, was an der Analogie gut funktioniert. Erstelle daraus anschließend einen One Pager oder einen Two Pager mit passenden fotorealistischen Bildern. Die Darstellung soll „Chez Gustav im Chaos“ mit der Situation nach der Umstellung der Terminierungslogik vergleichen.

ANMERKUNG ZUM GEGENENTWURF

Der applaudierende Schluss wurde später von einem der Modelle als zu idyllisch eingestuft, mit dem Argument, eine Geschäftsleitung lese ihn als Wunschbild. Der Einwand steht in Abschnitt 10. Er ist berechtigt und gleichzeitig eine Abwägung, die man auch anders treffen kann. Die Kontextdatei bleibt hier bewusst im Originalzustand, damit nachvollziehbar bleibt, worauf sich die Bewertungen beziehen.