

octonomy

DAS GROSSE BUCH DER VIRTUELLEN INGENIEURE

Solved.

Anfragewellen, die das Team nicht mehr treffen.
Expertenwissen, das in Sekunden auffindbar ist.
Antworten, auf die sich alle verlassen können —
Praxisbeispiele aus allen Branchen.

INHALT

Die Cases im Überblick

Einleitung — Warum die meisten KI-Projekte scheitern 03

Orientierung — Welcher Fall passt zu Ihrer Situation? 05

Der Kern — Der Context Layer 06

VERSICHERUNGEN

01 hepster 09

HANDEL

02 ANWR GROUP 18

03 celebrate company 26

MANUFACTURING

04 BORA 36

05 ALUCA 45

ENERGIE UND INFRASTRUKTUR

06 Prokon 51

LOGISTIK

07 Emons Spedition 57

SOFTWARE

08 ELAINE technologies 64

KLIMA- UND HEIZTECHNIK

09 Stiebel Eltron Schweiz 70

Was bleibt — Problem solved. Punkt? Nein, Komma. 73

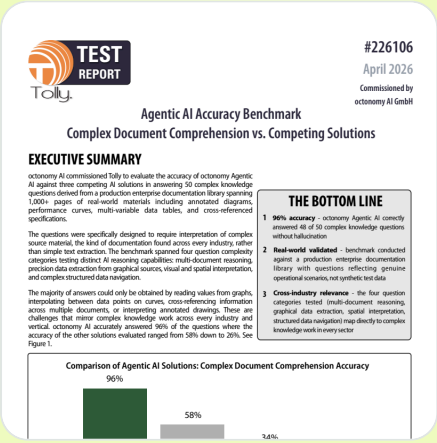
Die Plattform — Verstehen. Antworten. Handeln. 74

Warum die meisten KI-Projekte scheitern, und was diese Unternehmen anders gemacht haben

Die meisten KI-Projekte in Unternehmen halten nicht, was sie versprechen. 95 Prozent der GenAI-Piloten liefern keinen messbaren Mehrwert. Zwei Drittel aller intern entwickelten KI-Lösungen erreichen nie die Produktivumgebung. Der Anteil der Unternehmen, die den Großteil ihrer KI-Initiativen wieder einstellen, ist innerhalb eines Jahres von 17 auf 42 Prozent gestiegen.

Diese Zahlen stammen von MIT und S&P Global Market Intelligence aus dem Jahr 2025. Sie beschreiben ein Muster, das aus der Praxis vieler Unternehmen bekannt ist. In der Demo funktioniert KI. Im produktiven Betrieb, mit gewachsenen Systemen, echter Dokumentation und echten Kundenfragen, wird es schwierig.

Die Ursache ist in den meisten Fällen dieselbe. Unternehmen mit komplexem Wissen, technischen Produkten und umfangreicher Dokumentation kommen mit generischer KI nicht weit. Sie benötigen ein System, das ihr eigenes Wissen versteht, die passende Information findet und verlässlich damit antwortet. Die Antwort muss nachvollziehbar und frei von Halluzinationen sein und aus den Systemen kommen, die bereits im Einsatz sind.


TEST REPORT
Tolly.
#226106
April 2026
Commissioned by octonomy AI GmbH
Agentic AI Accuracy Benchmark
Complex Document Comprehension vs. Competing Solutions
EXECUTIVE SUMMARY
octonomy AI commissioned Tolly to evaluate the accuracy of octonomy Agentic AI against three competing AI solutions in answering 50 complex knowledge questions derived from a production enterprise documentation library spanning 1,000+ pages of real-world materials including annotated diagrams, performance curves, multi-variable data tables, and cross-referenced specifications.
The questions were specifically designed to require interpretation of complex source material, the kind of documentation found across every industry, rather than simple text extraction. The benchmark spanned four question complexity categories testing distinct AI reasoning capabilities: multi-document reasoning, precision data extraction from graphical sources, visual and spatial interpretation, and complex structured data navigation.
The majority of answers could only be obtained by reading values from graphs, interpolating between data points on curves, cross-referencing information across multiple documents, or interpreting annotated drawings. These are challenges that mirror complex knowledge work across every industry and vertical. octonomy AI accurately answered 96% of the questions where the accuracy of the other solutions evaluated ranged from 58% down to 26%. See Figure 1.
THE BOTTOM LINE
1 96% accuracy - octonomy Agentic AI correctly answered 48 of 50 complex knowledge questions without hallucination
2 Real-world validated - benchmark conducted against a production enterprise documentation library with questions reflecting genuine operational scenarios, not synthetic test data
3 Cross-industry relevance - the four question categories tested (multi-document reasoning, graphical data extraction, spatial interpretation, structured data navigation) map directly to complex knowledge work in every sector
Comparison of Agentic AI Solutions: Complex Document Comprehension Accuracy
96%
58%
26%

DIE STUDIE

Agentic AI Accuracy Benchmark: Complex Document Comprehension vs. Competing Solutions

Tolly Enterprises LLC, Test Report Nr. 226106, April 2026. Beauftragt von octonomy AI GmbH.

tolly.com/publications/226106 ↗

Wie groß der Abstand zwischen generischer und spezialisierter KI ausfällt, zeigt ein unabhängiger Test. Tolly Enterprises, ein Benchmarking-Institut mit mehr als 35 Jahren Erfahrung, hat octonomy im März 2026 gegen drei Wettbewerber getestet. Grundlage waren über 1.000 Seiten echter Industriedokumentation, darunter Schaltpläne und Leistungskurven. Bei einem Teil der Fragen verteilte sich die Antwort über mehrere Dokumente.

TOLLY BENCHMARK, MÄRZ 2026 · KORREKT BEANTWORTETE FRAGEN



octonomy beantwortete 48 von 50 Fragen korrekt. Der nächstbeste Wettbewerber arbeitete mit demselben zugrundeliegenden Sprachmodell. Quelle Marktdaten: MIT, „The GenAI Divide“, 2025. S&P Global Market Intelligence, 2025. Tolly Enterprises LLC, Benchmark Report Nr. 226106, März 2026.

Die Unternehmen in diesem Heft haben diesen Unterschied im laufenden Betrieb erfahren. Sie kommen aus unterschiedlichen Branchen und arbeiten mit unterschiedlichen Produkten und Teams. Ihre Ausgangslage war vergleichbar. Das Wissen lag im Unternehmen vor, ohne systematisch zugänglich zu sein. Der Support-Aufwand wuchs schneller als das Team. Routineanfragen banden Kapazität, die für die anspruchsvollen Fälle fehlte.

Alle haben sich für ein KI-System entschieden, das versteht, antwortet und handelt.

Dieses Heft erzählt ihre Geschichten, jede im gleichen Aufbau: Ausgangslage, Herausforderung, Ziele, Lösung, Umsetzung und Ergebnis, ergänzt um ein Interview mit dem Kunden und die Sicht des octonomy-Teams auf das Projekt. Alle Zahlen stammen von den Unternehmen selbst oder aus den Case Studies auf octonomy.ai. Nichts davon ist hinzugedichtet.

ORIENTIERUNG

Welcher Fall passt zu Ihrer Situation?

Die Kapitel sind nach Branchen sortiert. Die Einsatzfelder von octonomy verlaufen quer dazu. Grundlage dieser Übersicht ist deshalb nicht die Branche, sondern die Zielgruppe: Wer stellt die Frage, und welches Wissen wird für die Antwort gebraucht.

ENDKUNDEN

Kundenservice

Bestell-, Liefer- und Schadenstatus, Bedienhinweise, Reklamationen. Über Chat, E-Mail und Telefon, rund um die Uhr.

- ● hepster S. 09
- ● ANWR GROUP S. 18
- ● celebrate company S. 26
- ○ BORA S. 36

SUPPORT-TEAMS

Technischer Support

Komplexe Fälle mit Dokumentationstiefe: Handbücher, Schaltpläne, Fehlerlösungsbäume, Vorgangshistorie. Antwort in Sekunden, mit Quelle und Seitenzahl.

- ● BORA S. 36
- ○ ALUCA S. 45

TECHNIKER VOR ORT

Field Service

Diagnose vor der Anfahrt durch Verknüpfung von Fehlercodes und Monitoring-Daten mit der Herstellerdokumentation. Am Gerät dieselbe Quelle als Reparaturanweisung.

- ● ALUCA S. 45
- ● Prokon S. 51
- ● Stiebel Eltron Schweiz S. 70

FACHABTEILUNGEN

Interne Wissensaufbereitung

Regelwerke über tausende Seiten, IT-Dokumentation, Kundeninformationen in getrennten Systemen. Semantisch durchsuchbar, mit Kapitelangabe und Gesamtbild.

- ○ ANWR GROUP S. 18
- ● Emons Spedition S. 57
- ● ELAINE technologies S. 64

● Gefüllt: Schwerpunkt des Falls ○ Offen: im Kapitel ebenfalls behandelt

Jeder Fall trägt genau einen Schwerpunkt. Dessen Farbe kehrt auf der Auftaktseite des Falls wieder.

Der Context Layer: warum dieselbe Technologie bei jedem Kunden anders aussieht

Die Fälle in diesem Heft haben eine Gemeinsamkeit, die auf den ersten Blick nicht auffällt: Kein Kunde nutzt octonomy auf dieselbe Weise. hepster prüft Schadenfälle, BORA beantwortet Servicefragen zu Küchengeräten, Emons Spedition durchsucht Gefahrgut-Handbücher, Prokon diagnostiziert Windkraftanlagen. Dieselbe Technologie, unterschiedliche Anwendungen. Der gemeinsame Baustein dahinter ist der Context Layer.

Was der Context Layer ist

Moderne KI kann viel, aber sie weiß von sich aus nichts über Ihr Unternehmen. Sie kennt weder Ihre Produkte noch Ihre Regelwerke, Ihre Systeme oder die Art, wie bei Ihnen gearbeitet wird. Der Context Layer ist die Schicht, die genau das schließt: ein verstehender Kern zwischen Ihren Datenquellen und der Anwendung, die am Ende bei Ihnen läuft. An ihn lassen sich die unterschiedlichsten Quellen anschließen, und er führt zusammen, was für eine korrekte Antwort gebraucht wird.

Kontext ist dabei mehr als Daten. Damit ein KI-System im Arbeitsalltag verlässlich handelt, braucht es dreierlei:

Wissen

Wie Ihr Geschäft aufgebaut ist, also was ein Produkt, ein Vertrag oder ein Schadenfall bei Ihnen bedeutet.

Expertise

Wie die Arbeit tatsächlich abläuft, also die eingespielten Abläufe und Handgriffe, die oft nur erfahrene Mitarbeitende im Kopf haben.

Regeln

Was erlaubt ist, von Zuständigkeiten über Freigaben bis zu Datenschutzvorgaben.

Ein Teil davon steht in Handbüchern, ein großer Teil aber nur in gewachsenen Abläufen und in den Köpfen einzelner Menschen. Genau dieses verstreute Wissen führt der Context Layer zusammen.

Der Begriff wird inzwischen an vielen Stellen verwendet, meist aber für strukturierte Daten, für Datenbanken, Tabellen und Metadaten. Genau hier liegt der Unterschied bei octonomy: Unser Context Layer versteht auch statische, unstrukturierte Informationen, also das, was in Schaltplänen, technischen Zeichnungen, Handbüchern und annotierten Dokumenten steckt. Diese Fähigkeit, über unsere Visual-Cortex-Technologie, ist für viele unserer Kunden nicht ein Detail, sondern der eigentliche Grund, dass ihre Anwendung überhaupt funktioniert.

Das erklärt auch, woran die meisten KI-Projekte scheitern, von denen zu Beginn dieses Hefts die Rede war. Eine Demo überspringt diese Schicht oft und wirkt trotzdem beeindruckend. Im echten Betrieb rächt sich das: Ohne verlässlichen Context Layer beginnt ein Modell, Wissenslücken mit plausibel klingenden Erfindungen zu füllen, und die Antworten bleiben allgemein, statt zu Ihrem Fall zu passen. Der Aufbau dieser Schicht ist die eigentliche Arbeit, und genau sie unterscheidet ein System, das produktiv trägt, von einer Lösung, die es nur bis zur Vorführung schafft.

Warum das für Sie einen Unterschied macht

Ein häufiger Grund für das Scheitern von KI-Projekten liegt weniger in der Technologie als in der fehlenden Anpassung an das konkrete Unternehmen. Eine Standardlösung muss für viele Kunden gleichzeitig passen und bildet einen einzelnen Anwendungsfall daher nur begrenzt ab. Der Context Layer verschiebt diesen Ansatz: Statt ein Unternehmen an ein Standardprodukt anzupassen, wird die Anwendung auf den jeweiligen Anwendungsfall zugeschnitten.

Man kann sich das wie die Endstufe von Mass Customization vorstellen. Bei manchen Autoherstellern gleicht kein Fahrzeug dem anderen, weil sich über die Konfiguration so viel individualisieren lässt, dass am Ende genau Ihr Modell vom Band läuft. Bei Software war das lange nicht möglich, entweder man kaufte ein starres Standardprodukt oder man baute alles selbst. Der Context Layer verbindet beides: die Verlässlichkeit einer weiterentwickelten Plattform mit einer Anwendung, die sich anfühlt, als wäre sie nur für Sie gebaut.

Was octonomy dabei leistet

Die individuelle Anpassung bedeutet nicht, dass der Kunde die Lösung selbst warten muss. Sichtbar für den Kunden ist die Anwendung selbst: der jeweilige Anwendungsfall, die Sprache und das Erscheinungsbild des Unternehmens. Was darunter liegt, wird von octonomy gebaut und laufend weiterentwickelt, ebenso die Anbindung der Datenquellen und die laufende Wartung. Beim Kunden entsteht dadurch keine zusätzliche Wartungslast.

Aus diesem Aufbau ergibt sich, dass octonomy in den Fällen dieses Hefts meist nicht sichtbar in Erscheinung tritt. Die Anwendungen tragen den Namen und das Erscheinungsbild des jeweiligen Kunden. Die Plattform darunter ist in allen Fällen derselbe.



Versicherungen

In der Versicherungsbranche hat eine falsche Antwort direkte finanzielle und rechtliche Folgen. Allgemeine Versicherungsbedingungen sind rechtlich präzise, kontextabhängig und produktspezifisch. Eine korrekte Antwort setzt voraus, dass Klauseln, Ausschlüsse und Vertragsstatus zugleich verstanden werden. Die Schadensbearbeitung verläuft über mehrere Stufen und erfordert Systemzugriffe, Regellogik und Dokumentation. Ein Chatbot allein leistet das nicht. Mit wachsender Produktvielfalt steigt das Anfragevolumen schneller als das Team, und jede Antwort muss regulatorisch und intern nachvollziehbar bleiben. Genau hier setzt octonomy an, mit vollständigem Verständnis von AVB, Policen und Vertragsdaten, mit geprüfter Quellenreferenz statt Halluzinationen und mit Vorgängen, die end-to-end abgeschlossen werden.

01 · hepster zeigt in diesem Kapitel, wie das in der Praxis funktioniert.

hepster

hepster

>50 %	0	linear
Reduktion bei Schadenbearbeitung und Support	Qualitätseinbußen	fängt Teamskalierung bei Unternehmenswachstum ab



„Dadurch, dass wir diese Last weggenommen haben, haben wir viel mehr Zeit für die Menschen, die die Zeit mit uns brauchen."

Katja Fröhlich, Tech Leader IT & Digital Products, hepster

AUF EINEN BLICK

Das Unternehmen	InsurTech aus Rostock mit Fokus auf Embedded Insurance in Mobility, Electronics, Sports, Pets, Travel und weiteren Bereichen. Eingesetzt im Kundenservice und in der Schadenprüfung.
Ausgangslage	Mit dem Produkt- und Partnergeschäft wächst die Vielfalt der Anfragen. Damit steigt der Druck auf den Kundenservice, der schnelle Reaktionszeiten, korrekte Antworten und eine saubere Dokumentation liefern muss.
Was hepster gemacht hat	hepster ergänzt den Kundenservice um die octonomy KI-Expertenlösung. Sie versteht Anfragen, nutzt Informationen aus Systemen und Wissensquellen und schließt Prozesse im Idealfall end-to-end ab.
Ergebnis	Der Aufwand in Schadenbearbeitung und Kundensupport sank um 50 Prozent, ohne Qualitätseinbußen.

VIDEOS UND MATERIAL

CASE STUDY

WEBINAR

Die Case Study

AUSGANGSLAGE

hepster ist ein InsurTech mit Fokus auf Embedded Insurance, also auf Versicherungslösungen, die sich nahtlos in digitale Customer Journeys integrieren. Das Angebot reicht unter anderem in die Bereiche Mobility, Electronics, Sports, Pets, Travel, Furniture und Luxury Goods. Für Partner setzt hepster auf API-basierte Integration, damit Versicherungen schnell Teil bestehender Prozesse werden.

Mit dem Produkt- und Partnergeschäft wächst die Vielfalt der Anfragen. Damit steigt der Druck auf den Kundenservice, der schnelle Reaktionszeiten, korrekte Antworten und eine saubere Dokumentation liefern muss.

HERAUSFORDERUNG

Das Embedded-Insurance-Geschäft bringt viele Kontexte, viele Regeln und viele Sonderfälle mit sich, von Produktdetails über Vertrags- und Tariflogik bis zu Dokumenten und Nachweisen. Zugleich erwarten Kundinnen und Kunden schnelle, präzise Antworten über digitale Kanäle. Klassische Chatbots helfen hier laut hepster nur begrenzt, weil ihnen der Systemzugriff, die belastbare Nachvollziehbarkeit und die echte Prozessbearbeitung fehlen. Diese Lücke war der Startpunkt. Ziel war eine Automatisierung, bei der Qualität, Compliance und Kundenerlebnis nicht leiden.

ZIELE

hepster hat den Fokus auf drei Ziele gelegt.

Entlastung im Tagesgeschäft

Weniger manuelle Routinearbeit, mehr Zeit für komplexe Fälle.

Konstante Antwortqualität

Verlässliche Aussagen statt Best-Effort-Antworten eines klassischen Chatbots.

Skalierbarkeit

Neue Themen, Produkte und Partner sollen den Support nicht linear vergrößern.

LÖSUNG

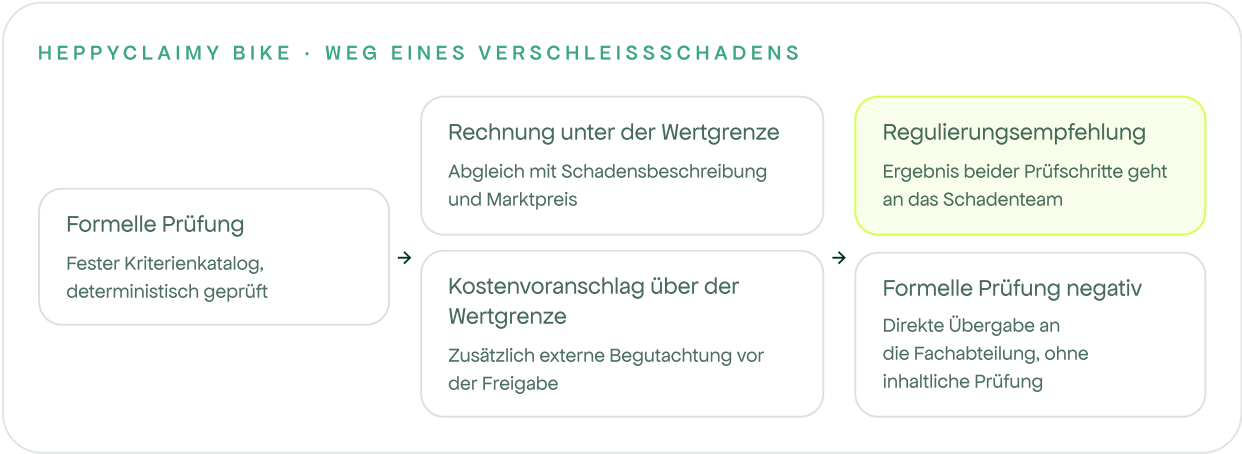
octonomy arbeitet als Multi-Agenten-System und damit als koordiniertes Team spezialisierter Agenten, etwa für Vertragsfragen, Reklamationen, Zahlungen und Dokumente, ergänzt um einen Supervisor- und einen Ops-Agent für Planung, Routing und Dokumentation. Entscheidend für den Support-Alltag ist der End-to-End-Ansatz. Anliegen werden verstanden, relevante Informationen auch in Dokumenten geprüft und der Vorgang sauber abgeschlossen, inklusive Rückmeldung an die Kunden. Die Anbindung an CRM und Ticketsysteme läuft über APIs und Konnektoren, die Anbindung an Policy- und Billing-Systeme ist geplant. Schritte und Entscheidungen sind protokollierbar, und die Antworten basieren strikt auf geprüften Wissensquellen. Die Lösung ist laut hepster mehrsprachig und skalierbar, die Verarbeitung erfolgt DSGVO-orientiert nach europäischen Standards.

UMSETZUNG

hepster ist pragmatisch gestartet. Zunächst wurden ein bis zwei klar umrissene Use Cases definiert, KPIs wie Entlastung, Antwortzeit und Lösungsquote festgelegt, der Pilot live gestellt und iterativ erweitert. In der ersten Stufe liefen die Kanäle E-Mail, Chat und Kontaktformular, auf Basis von Ticketsystem, CMS und einer Wissensbasis mit AVB, Produktunterlagen, internen Richtlinien und Prozessdokumenten. Einfache Chat-Interaktionen und Routinefälle wurden automatisiert beantwortet, komplexe Fälle mit Kontext an Menschen übergeben. In der zweiten Ausbaustufe kamen das Kundenkonto und die Anbindung an das Policy-System hinzu, mit Direktantworten auf Ticketanfragen und der Umsetzung von Standardanfragen wie Adress- oder Namensänderungen über einen Admin-Ops-Agenten.

ERGEBNIS

Der Aufwand in Schadenbearbeitung und Kundensupport sank um 50 Prozent, ohne Qualitätseinbußen. Über den skalierbaren Automatisierungs-Layer lassen sich weitere Prozesse Schritt für Schritt ergänzen, etwa Statusabfragen, Stammdaten- und Vertragsänderungen sowie die generelle Informationsbereitstellung. Gerade in Phasen von Wachstum und neuen Partnerschaften bleibt laut hepster das Ziel bestehen, den Support ohne lineares Teamwachstum zu skalieren.



Our Side — Wie wir es gebaut haben

Ausgangslage: zwei Sparten, ein gemeinsamer Engpass

hepster wächst in mehreren Sparten, unter anderem in der Tierversicherung und der Fahrradversicherung. Mit dem Wachstum stieg das Schadenvolumen und damit der Anteil repetitiver Prüfarbeit. Das Erfassen eines Schadens war dabei selten das Problem. Der eigentliche Aufwand entstand bei der Prüfung. Bevor eine Rechnung oder ein Kostenvoranschlag reguliert werden konnte, musste beides gegen Schadensbeschreibung und Versicherungsbedingungen abgeglichen werden, und die dafür nötigen Informationen lagen über Partnerportal, Ticketsystem, Zahlungsdienstleister und Reporting-Tool verteilt. Für die Tierversicherung liefert die Gebührenordnung für Tierärzte ein einheitliches Regelwerk. Für Fahrräder existiert kein solches Regelwerk, weil jeder Fachhändler eine eigene Rechnungslogik und eigene Arbeitswerte ansetzt.

Für automatisierte Entscheidungen brauchte es eine höhere Präzision, als die bis dahin eingesetzten Mittel liefern konnten.

Architekturprinzip: formell und inhaltlich sauber getrennt

Alle drei Assistenten laufen auf der octonomy-Plattform und sind über das Ticketsystem an den Arbeitsalltag der Fachabteilung angebunden. Löst ein Ticket einen Assistenten aus, erhält dieser zunächst nur die Ticket-ID. Alle weiteren Informationen beschafft er selbst, die Vertragsdaten aus dem Partnerportal, die Zahlungsverläufe über den Zahlungsdienstleister und die Dokumente aus der Cloud-Ablage. Am Ende schreibt der Assistent sein Ergebnis als internen Kommentar zurück ins Ticket, für den Kunden unsichtbar.

Beiden Schadenassistenten liegt dieselbe Grundidee zugrunde, eine klare Trennung zwischen zwei Prüfschritten.



PRÜFSCHRITT	LOGIK	BEISPIELE
Formelle Prüfung	Deterministisch umgesetzt, nach festen und eindeutigen Regeln statt über ein Sprachmodell	Ist der Vertrag aktiv? Wurde eine Wartezeit eingehalten?
Inhaltliche (materielle) Prüfung	Beginnt erst nach positiver formeller Prüfung. Hier gleicht das Sprachmodell Rechnung oder Kostenvoranschlag mit Schadensbeschreibung und Versicherungsbedingungen ab	Passt die Rechnung zum gemeldeten Schaden? Ist der Preis plausibel?

Diese Trennung war keine Formsache. Alles, was sich eindeutig regelbasiert entscheiden lässt, läuft als eigenständige, deterministische Logik und erreicht das Sprachmodell bewusst nicht. Auf diese Weise konzentriert sich das Sprachmodell auf die Aufgabe, für die es tatsächlich gebraucht wird, den inhaltlichen Abgleich, bei dem Kontext und Bewertung gefragt sind.

Die drei Assistenten im Überblick

ASSISTENT	ZIELGRUPPE	AUFGABE
heppyChat	Endkunden	Beantwortet allgemeine Fragen zu Produkten, Verträgen und Prozessen
heppyClaimy PET	Schadenabteilung (Tierversicherung)	Prüft Schadenfälle und erstellt eine Regulierungsempfehlung zur weiteren Bearbeitung durch das Schadenteam
heppyClaimy Bike	Schadenabteilung (Fahrradversicherung, Verschleiß)	Formelle und inhaltliche Prüfung von Verschleißschäden

heppyChat ist der Kundenservice-Chatbot und beantwortet allgemeine Fragen von Endkunden zu Produkten, Verträgen und Prozessen. Er stützt sich dabei auf die bestehende Wissensdatenbank und ist produktiv im Einsatz.

heppyClaimy PET unterstützt die Schadenabteilung bei Tierversicherungsschäden. Sobald ein Fall die formelle Prüfung durchlaufen hat, gleicht der Assistent Rechnung und Diagnose gegen die Gebührenordnung für Tierärzte, die Vertragsbedingungen und eine Liste erstattungsfähiger Medikamente ab. Am Ende steht eine Regulierungsempfehlung, die das Schadenteam weiterbearbeitet. Auch dieser Assistent ist produktiv.

heppyClaimy Bike übernimmt dieselbe Aufgabe für Fahrradversicherungen und ist aktuell auf Verschleißschäden begrenzt. Die formelle Prüfung deckt mehrere vertragliche und zeitliche Kriterien ab und läuft bereits produktiv. Für die inhaltliche Prüfung gibt es zwei Wege. Rechnungen unterhalb einer definierten Wertgrenze werden direkt gegen Schadensbeschreibung und Marktpreis geprüft. Kostenvoranschläge darüber durchlaufen vor der Freigabe zusätzlich eine externe Begutachtung.

Die Ersatzteilpreise ermittelt der Assistent durch gezielte Recherche direkt bei den Herstellern, die Arbeitszeiten prüft er anhand einer branchenüblichen Referenzliste. Das Ergebnis beider Prüfschritte geht als Empfehlung an die Fachabteilung, die über die Regulierung entscheidet. Die formelle Prüfung ist produktiv, die inhaltliche Prüfung wird schrittweise ausgerollt.

Wie wir Vertrauen aufgebaut haben

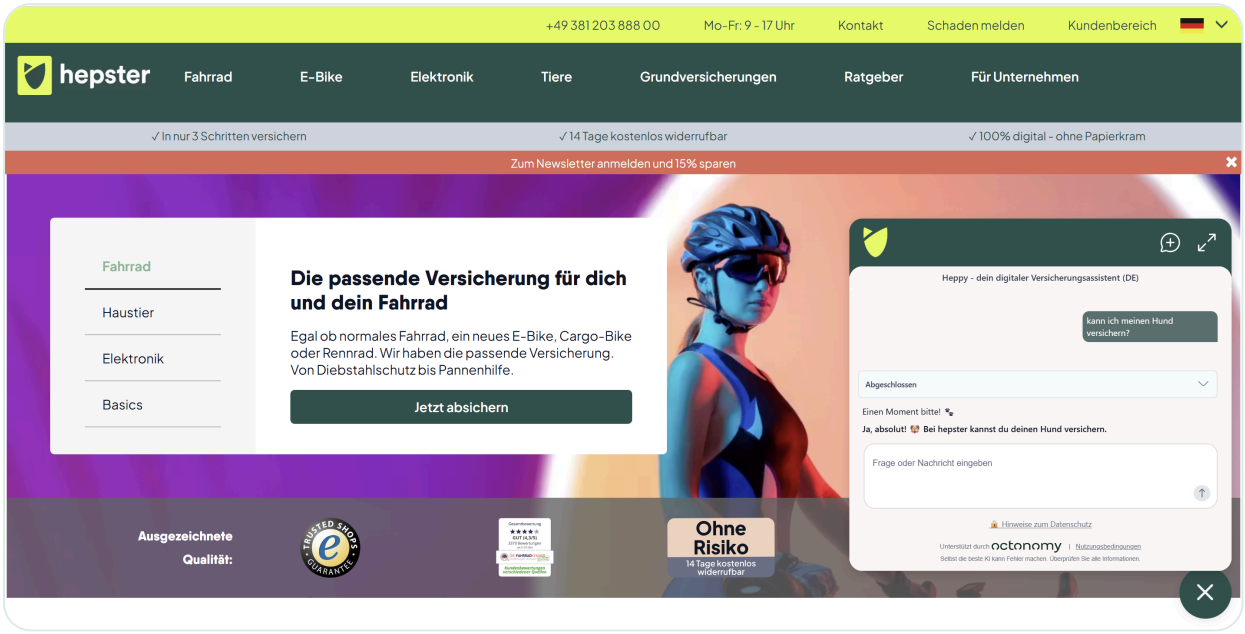
Zahlen allein erzählen nicht die ganze Geschichte. heppyClaimy PET bearbeitet inzwischen eine vierstellige Fallzahl pro Monat und liefert eine sehr hohe Trefferquote bei den verwertbaren Empfehlungen. Für die Belastbarkeit dieser Werte sorgt ein fester Testrahmen. hepster dokumentiert für jede der vier möglichen Entscheidungen reale Schadenfälle mit Ticketnummer, die als Referenz für Kalibrierung und laufende Qualitätssicherung dienen. Angenommene Beispiele kommen dabei nicht zum Einsatz. Zur laufenden Überwachung nutzen wir zusätzlich ein Feature, das wiederkehrende Fehlermuster automatisch erkennt, sobald sich vergleichbare Fälle innerhalb kurzer Zeit häufen.

DER TESTRAHMEN HINTER HEPPYCLAIMY PET

Referenzfälle	Mustererkennung	Im Betrieb
Je Entscheidung ein realer Schadenfall mit Ticketnummer, keine angenommenen Beispiele	Alarm ab fünf vergleichbaren Fehlerfällen je Woche, etwa 19 Fälle ohne Chipnummer	Vierstellige Fallzahl pro Monat, sehr hohe Quote verwertbarer Empfehlungen

heppyChat im Einsatz

So erreichen Endkundinnen und Endkunden den Kundenservice-Assistenten auf der hepster-Website. Er antwortet in natürlicher Sprache, direkt aus der freigegebenen Wissensbasis, und liefert eine Tarifübersicht mit weiterführendem Link, statt ins Kontaktformular weiterzuleiten.



Screenshot: hepster.com

Antwort statt Verweis

Der Assistent nennt Produkte und Tarife konkret, inklusive Selbstbeteiligung und Jahreshöchstgrenze.

Auf geprüftem Wissen

Grundlage sind ausschließlich freigegebene hepster-Inhalte. Fehlt eine Information, wird das offen gesagt.

Transparent gekennzeichnet

Der Hinweis auf die KI-Generierung und die Datenschutzangaben sind Teil der Oberfläche.

Customer Side — Interview mit hepster

Auszug aus dem Gespräch mit Martin Lichtwark, Lead Data & AI, und Katja Fröhlich, Tech Leader IT & Digital Products, hepster.

Der Begriff Chatbot war intern eigentlich verbrannt. Was hat euch trotzdem überzeugt?

> „Der Begriff Chatbot war bei uns intern sehr verbrannt. Uns hat einfach die Leistungsfähigkeit überzeugt, weil wir sehr spezifisches Wissen und sehr spezifische Fragen zu unseren Produkten haben, sodass ein Agentenansatz eine Möglichkeit war, gezielt zu antworten, genau in dem Spektrum, das wir für unsere Kunden brauchen."

Martin Lichtwark, Head of Data & AI, hepster

Warum habt ihr ausgerechnet mit Produktinformationen zu Versicherungen gestartet?

> „Wir hatten einen ganz klaren Auftrag: Produktinformation vernünftig zu vermitteln. Der Teufel steckt immer im Detail, und es ist umso wichtiger, dass wir am Anfang die richtigen Informationen geben, auch aus dem Kleingedruckten, damit der Kunde weiß, dass das, was er dort kauft, auch das ist, was er im Notfall braucht."

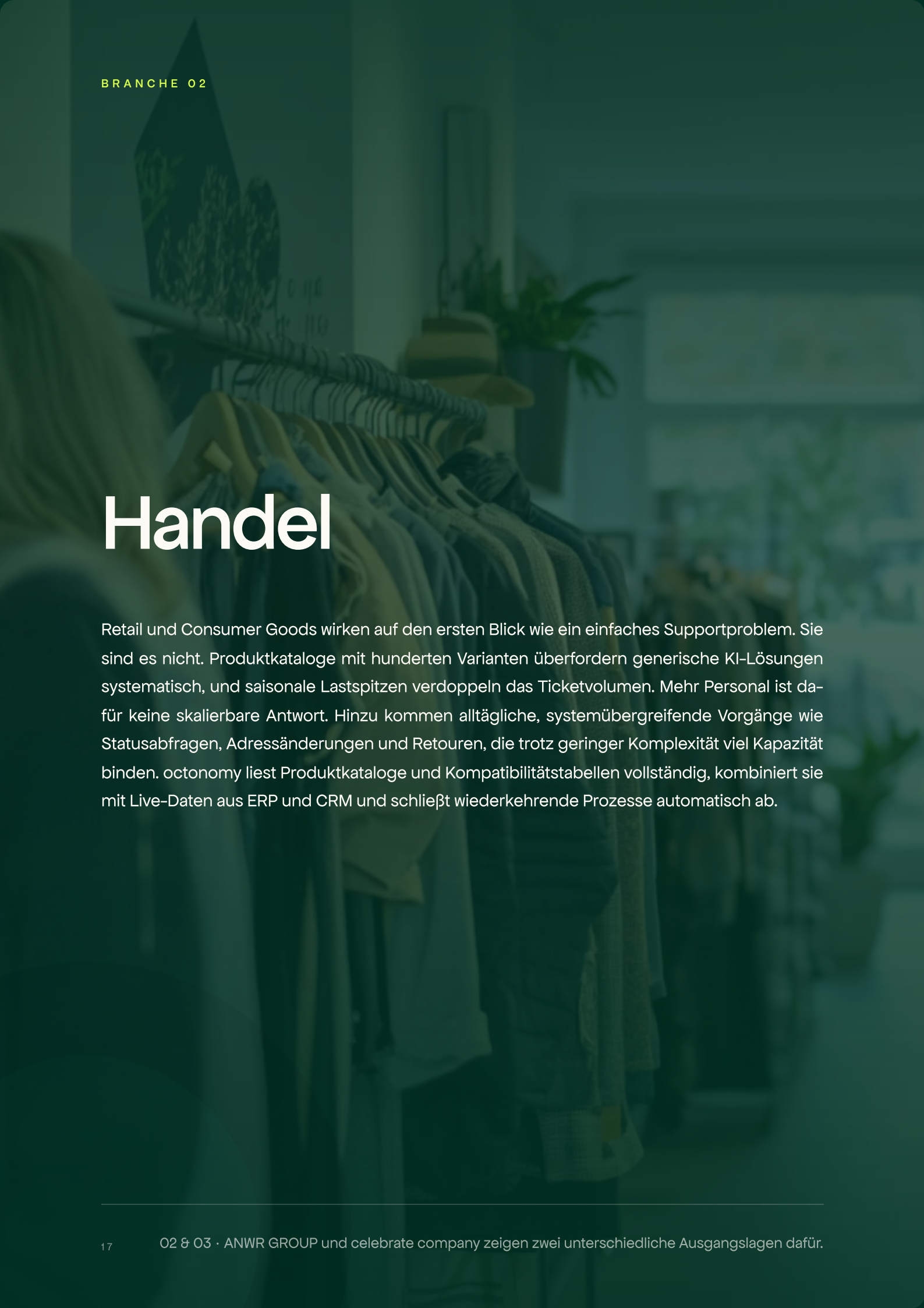
Katja Fröhlich, Tech Leader IT & Digital Products, hepster

Was hat sich für euer Team im Alltag konkret verändert?

> „Diese stumpfe Routinearbeit ist deutlich weniger geworden. Die Menschen stehen nicht mehr so sehr unter dem operativen Druck, sie haben die Zeit und auch die Nerven, sich um die komplexen Themen und Menschen zu kümmern, die einfach ein bisschen mehr Aufmerksamkeit brauchen."

Katja Fröhlich, Tech Leader IT & Digital Products, hepster

Support skaliert, wenn das System das Volumen trägt, nicht das Team.



BRANCHE 02

Handel

Retail und Consumer Goods wirken auf den ersten Blick wie ein einfaches Supportproblem. Sie sind es nicht. Produktkataloge mit hunderten Varianten überfordern generische KI-Lösungen systematisch, und saisonale Lastspitzen verdoppeln das Ticketvolumen. Mehr Personal ist dafür keine skalierbare Antwort. Hinzu kommen alltägliche, systemübergreifende Vorgänge wie Statusabfragen, Adressänderungen und Retouren, die trotz geringer Komplexität viel Kapazität binden. octonomy liest Produktkataloge und Kompatibilitätstabellen vollständig, kombiniert sie mit Live-Daten aus ERP und CRM und schließt wiederkehrende Prozesse automatisch ab.

02 & 03 · ANWR GROUP und celebrate company zeigen zwei unterschiedliche Ausgangslagen dafür.

ANWR GROUP

75 %	~1/3	20 %	6–9 Mon.
Direktlösung im Chat	Ticket-Automatisierung	Händler-Nutzung ANNI	ROI



„Wir können sagen, dass wir im Chatbot ca. 75 Prozent aller Anfragen direkt im Chatbot lösen können."

Sven Kulikowsky, CIO, ANWR GROUP

AUF EINEN BLICK

Das Unternehmen	Handelskooperation für Schuh-, Sport- und Lederwarenhandel in Europa, gegründet 1919, mit 4.300 angeschlossenen Händlern. Eingesetzt in Händlerberatung, E-Commerce-Service und interner IT.
Ausgangslage	Viele Verbundgesellschaften mit einem über Jahrzehnte gewachsenen Serviceangebot. Für Händler war oft unübersichtlich, welchen Service es gibt und wer zuständig ist. Im E-Commerce häuften sich repetitive Kundenanfragen.
Was ANWR gemacht hat	Drei Use Cases mit octonomy: ANNI unterstützt die Handelspartner, TESS den internen IT-Support. Hinzu kommen ein Chatbot und ein OpsAgent für den E-Commerce, die Endkunden rund um die Uhr bedienen und Tickets automatisiert bearbeiten.
Ergebnis	20 % der angeschlossenen Händler nutzen ANNI bereits aktiv. Im E-Commerce löst der Chatbot 75 % aller Anfragen direkt im Chat. Rund ein Drittel aller eingehenden Tickets bearbeitet der OpsAgent vollautomatisch. Den ROI erreichte die ANWR GROUP innerhalb von sechs bis neun Monaten.

VIDEOS UND MATERIAL

CASE STUDY

MASTERCLASS

SUCCESS STORY

Die Case Study

AUSGANGSLAGE

Die ANWR GROUP zählt zu den erfolgreichsten und umsatzstärksten Handelskooperationen in Europa, mit Schwerpunkt auf dem Schuh-, Sport- und Lederwarenhandel. Gegründet wurde sie 1919, ihr Hauptsitz liegt in Mainhausen, und sie ist in 19 europäischen Ländern tätig. Zum Verbund gehören 4.300 angeschlossene Händler mit rund 9.000 Filialen, 1.400 Mitarbeitende und 2.500 ZR-Lieferanten. Die Gruppe vereint Handels- und Kooperationsgesellschaften wie ANWR Schuh, SPORT 2000 und Goldkrone unter einem Dach, dazu die beiden Banken DZB BANK und AKTIVBANK sowie Immobilien- und Beteiligungsgesellschaften.

Die ANWR GROUP besteht aus vielen Verbundgesellschaften mit einem breiten, über Jahrzehnte gewachsenen Serviceangebot. Für Händler war es oft unübersichtlich, welchen Service es gibt, wer zuständig ist und wie sie ihn erreichen. Im E-Commerce häuften sich repetitive Kundenanfragen, die Mitarbeitende täglich dutzende Male beantworteten.

HERAUSFORDERUNG

Über 100 Jahre Unternehmensgeschichte haben ein breites und wertvolles, zugleich aber komplexes Serviceangebot entstehen lassen, mit vielen Verbundgesellschaften, vielen Ansprechpartnern und vielen Prozessen. Für angeschlossene Händler, überwiegend kleine und mittelständische Betriebe, ist es schwer zu navigieren. Welchen Service gibt es, wer ist zuständig, und wie erreiche ich ihn.

> „Für unsere Händler war es schwer zu verstehen, wo kriege ich welchen Service her. Wir wollten wissen: Wie können wir unseren Kunden unseren Service einfacher zur Verfügung stellen?“
>
> Sven Kulikowsky, CIO, ANWR GROUP

Gleichzeitig betreibt die ANWR GROUP E-Commerce-Portale für ihre Schuh-, Sport- und Lederwarenhändler, darunter schuhe.de, sport2000.de und bagmondo.de. Bis zu 70 Prozent der eingehenden Kundenanfragen dort sind repetitiv und drehen sich um den Verbleib des Pakets, den Ablauf der Retoure oder den passenden Aufkleber. In Peak-Phasen wie Black Friday waren die Service Level kaum einzuhalten, und nach Feierabend gab es keine Antworten.

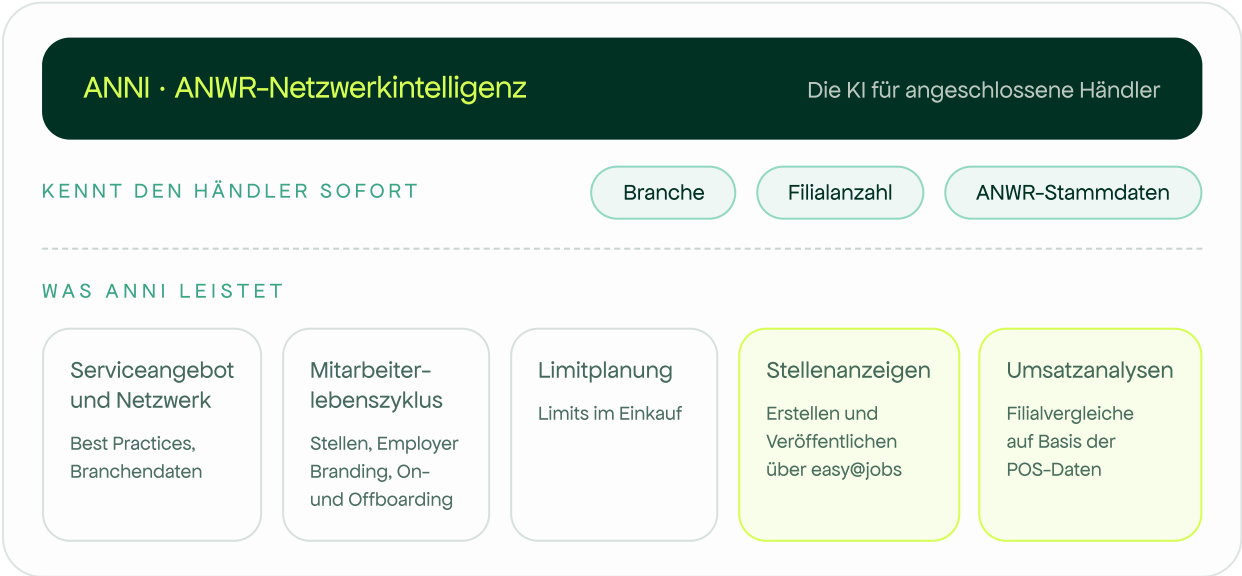
ZIELE

Drei Ziele standen im Fokus. Das erste war eine schnelle und personalisierte Händlerberatung, mit Antworten auf Knopfdruck, individuell auf das eigene Unternehmen zugeschnitten und ohne lange Suche nach dem richtigen Ansprechpartner. Das zweite war die Auslösung von Prozessen direkt aus dem Chat, also Stellenanzeigen erstellen und veröffentlichen, Umsatzanalysen abrufen und Bestellungen auslösen, ohne das Chat-Fenster zu verlassen. Das dritte war die Automatisierung des Endkundenservice, damit sich Mitarbeitende auf qualitative Themen konzentrieren können und Kunden auch am Wochenende eine Antwort erhalten.

LÖSUNG

ANNI, die ANWR-Netzwerkintelligenz, ist die KI für angeschlossene Händler. Sie beantwortet Fragen rund um das ANWR-Serviceangebot, teilt Best Practices aus dem Netzwerk, liefert Branchendaten und gibt personalisierte Antworten auf Basis der eigenen Unternehmensdaten des Händlers, direkt über die Schnittstelle zu den ANWR-Stammdaten. Sobald ein Händler den Chat öffnet, weiß ANNI bereits, ob es sich um einen Schuh-, Sport- oder Lederwarenhändler handelt und wie viele Filialen er betreibt. Zu den konkreten Funktionen zählen die Beratung entlang des Mitarbeiterlebenszyklus, etwa Stellenausschreibungen, Employer Branding sowie Onboarding- und Offboarding-Checklisten, dazu Themen wie Limitplanung und Einkaufslimits, das Erstellen und Veröffentlichen von Stellenanzeigen über easy@jobs sowie Umsatzanalysen und Filialvergleiche auf Basis angebundener POS-Daten.

Für den E-Commerce arbeiten zwei Komponenten zusammen. Ein Chatbot übernimmt den Customer Self Service direkt auf den E-Commerce-Seiten. Der OpsAgent, intern E-Mail-Türsteher genannt, liest eingehende Tickets aus dem CRM SAP C4C, bewertet, welche er selbst lösen kann, und beantwortet diese vollständig, inklusive Retourendokumenten, Storno-Informationen und Lieferstatus direkt über die DHL-API. Was er nicht lösen kann, leitet er mit Kontext an Mitarbeitende weiter. Angebunden sind C4C für Stammdaten und Tickets, Qualibet für Bestellinformationen, Confluence für Prozessdokumentation und DHL für Live-Tracking.



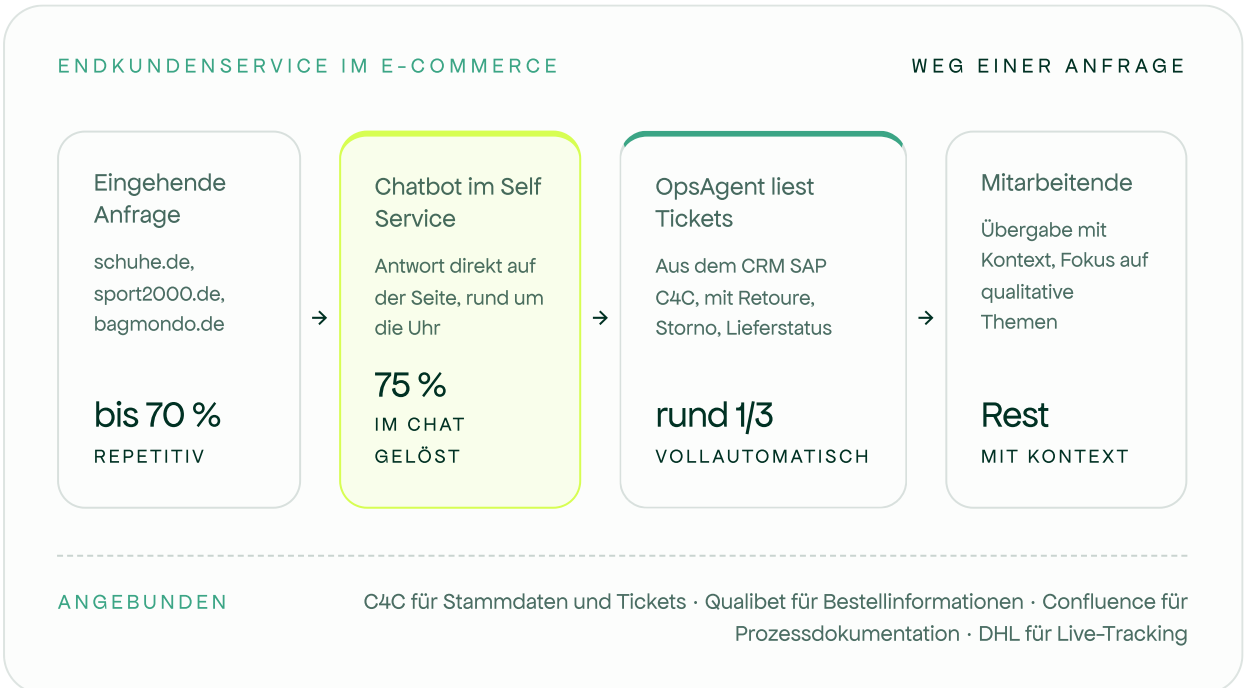
UMSETZUNG

Die ANWR GROUP und octonomy sind pragmatisch gestartet. Von der Entscheidung bis zum Go-Live vergingen drei Monate. Der erste ANNI-Launch war an ein Händlerevent gebunden, bei dem der Vorstand das Thema auf der Bühne präsentierte, und diese harte Deadline hat Geschwindigkeit erzeugt. Das Projekt wurde bewusst als gemeinsames Vorhaben von IT und Fachbereich aufgesetzt, mit kurzen Entscheidungswegen und enger Abstimmung, statt als IT-Projekt mit nachgelagerter fachlicher Abnahme. Den Start bildete ein klar abgegrenzter POC zum Thema Mitarbeiter binden und finden, darauf aufbauend kamen sukzessive weitere Themen, Schnittstellen und Use Cases hinzu. Heute laufen ANNI, Chatbot, OpsAgent und künftige Anwendungen auf derselben technischen Basis über octonomy, dem ANWR AI Hub.

> „Die Mitarbeiter werden entlastet. Diese profanen Tickets werden weggenommen, und insofern haben die Mitarbeiter viel mehr Zeit für qualitative Themen."
>
> Sven Kulikowsky, CIO, ANWR GROUP

ERGEBNIS

75 Prozent aller Chatbot-Anfragen löst der Agent direkt im Chat, ohne dass ein Mitarbeiter eingreift. Rund ein Drittel aller eingehenden Tickets bearbeitet der OpsAgent vollautomatisch. 20 Prozent der angeschlossenen Händler nutzen ANNI bereits aktiv. Den ROI erreichte die ANWR GROUP innerhalb von sechs bis neun Monaten. Das langfristige Ziel ist, dass Händler ANNI als einzigen Einstiegspunkt nutzen, um Informationen abzurufen, Prozesse anzustoßen und Ware zu bestellen, ohne zwischen Portalen zu wechseln. Im E-Commerce soll die Automatisierungsrate weiter steigen, mit dem Ziel, von reinem Service zu aktiver Kundenberatung zu wachsen.



Our Side — Wie wir es gebaut haben

octonomy hat für die ANWR GROUP keinen einzelnen Chatbot gebaut, sondern drei eigenständige KI-Assistenten für drei sehr unterschiedliche Zielgruppen: Händler, Mitarbeitende und Endkunden. Alle drei laufen auf derselben octonomy-Plattform. Jeder ist für seine Zielgruppe eigenständig gedacht.

Die ANWR GROUP hat 4.300 Handelspartner, 1.400 Mitarbeitende und dazu Endkunden, die auch nachts um elf wissen wollen, wo ihre Schuhe bleiben. Drei sehr unterschiedliche Zielgruppen, drei KI-Assistenten. Dieser Abschnitt zeigt, wie octonomy sie aufgebaut hat.

ANNI

HANDELSPARTNER

Zentraler Chat-Einstieg über Schuh, Sport 2000 und Goldkrone

TESS

MITARBEITENDE

IT-Support in der bestehenden Hilfeseite

Order-Support

ENDKUNDEN

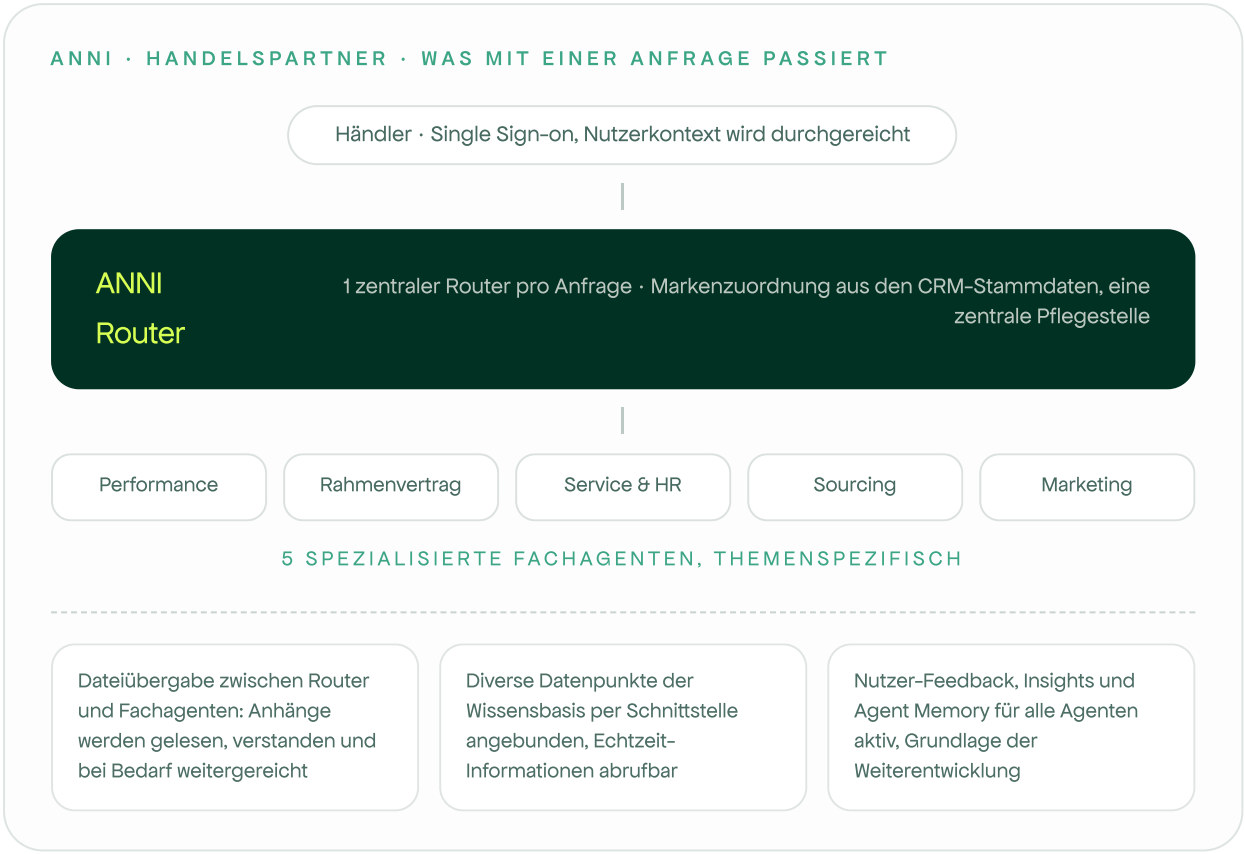
Bestellstatus, Retouren und Rechnungen auf schuhe.de und Sport 2000

ANNI: der Kompass für den Handelspartner

Ein Router, der Anfragen richtig verteilt, ist nur die Grundlage. Der eigentliche Nutzen entsteht einen Schritt später. Er entsteht dort, wo ein Fachagent rund um die Uhr eine Frage beantwortet, die ein Händler sonst nur mit einem Anruf im Innendienst und außerhalb dessen Erreichbarkeit gar nicht geklärt bekäme.

Aus dieser Überlegung ist ANNI entstanden. ANNI ist der zentrale Chat-Einstieg für die Handelspartner der ANWR GROUP, über die Marken Schuh, Sport 2000 und Goldkrone hinweg. Fachlich berät ANNI nicht selbst. Sie ordnet jeden Händler nach der einmaligen Anmeldung per Single Sign-on der richtigen Marke zu und gibt die Anfrage im Hintergrund an den passenden Fachagenten weiter. Der Händler merkt davon nichts: Er führt einfach ein Gespräch, das die richtige Antwort liefert, ohne zu wissen, mit wie vielen Agenten er dabei verbunden ist.

Die Fragen, die auf diesem Weg beantwortet werden, sind entsprechend vielfältig: Wie lief meine Saison im Vergleich zu vergleichbaren Häusern? Was verkauft sich gerade wirklich? Was sagt der Rahmenvertrag zu diesem Fall? Wie erstelle ich eine Stellenausschreibung? Welche Marketingmaßnahmen eignen sich für mein geplantes Event? Möglich macht diese Bandbreite eine Multi-Agenten-Architektur anstelle eines einzelnen, monolithischen Systems.



Das Ziel ist ein zentraler Einstieg für die Händler. Eine getrennte Anmeldung je nach Thema entfällt. Zudem kann die Wissensbasis kontinuierlich erweitert werden. Für den Händler ändert sich nichts, außer dass Neuerungen ihn ab sofort zeitgleich mit allen anderen erreichen.

Den größten Sprung hat der Performance-Assistent im vergangenen Jahr gemacht. Die Vor-Order-Planung rechnet saisonale Zeiträume durch. Die Bestsellerlisten zeigen, was sich tatsächlich bewegt. Der Clustervergleich setzt beides in Bezug zu vergleichbaren Händlern. Früher gab es dafür einen Bericht mit 68 Filtern, und schon dessen Bedienung war eine Hürde. Mit ANNI genügt eine einzige Frage im Chat, und der Händler bekommt seine Auswertung.

Auch die Frage nach der Stellenausschreibung landet nicht im Leeren. Dahinter steckt der Sourcing-Agent, der Händlern hilft, Stellenanzeigen für ihre eigenen Läden zu formulieren und zu veröffentlichen. Das ist Recruiting-Unterstützung an einer Stelle, an der sonst oft die Zeit oder die Vorlage fehlt.

TESS: der IT-Kollege, der nie schläft

TESS arbeitet nach innen, ANNI nach außen. TESS richtet sich an die Mitarbeitenden der ANWR GROUP. Bisher sortiert der Service Desk 30 bis 50 IT-Anfragen am Tag von Hand vor, über ein Formular und ein Postfach. TESS übernimmt diese Vorarbeit, eingebettet als vertrautes Fenster direkt in der bestehenden IT-Hilfeseite.

Der Anspruch dabei ist es, möglichst selten ein Ticket schreiben zu müssen. TESS durchsucht zuerst die interne Wissensdatenbank, danach das Web. Dabei erarbeitet sie mindestens drei unterschiedliche Lösungsansätze, bevor ein Ticket überhaupt zur Option wird. Erst wenn das nicht reicht, ordnet sie den Fall als Störung oder Aufgabe ein. Dann übergibt sie ihn strukturiert an das Ticketsystem, mit vollständigem, wortgetreuem Chatverlauf.



Rund um die Uhr für den Endkunden

Der dritte Assistent richtet sich an die Endkunden von schuhe.de und sport2000.de. Der Order-Support ist auf den Webseiten eingebunden und arbeitet direkt im Bestellprozess. Fragen zu Bestellstatus, Rücksendungen und Rechnungen beantwortet er sofort über die bestehende Bestell-Schnittstelle, ohne Warteschleife, auch um elf Uhr abends. Einen Vorgang im Ticketsystem legt er nur an, wenn es wirklich einen Menschen braucht.

Dahinter arbeitet der Ops-Agent dem Support-Team zu, unsichtbar für den Endkunden. Er ordnet eingehende Fälle ein und prüft, ob er sie vollautomatisch beantworten kann. Wenn nicht, steuert er den Fall an das Support-Team aus. Das geschieht regelbasiert und auf Grundlage der starken Wissensbasis. So wird das Support-Team im Tagesgeschäft spürbar entlastet.

Customer Side — Interview mit ANWR GROUP

Auszug aus dem Gespräch mit Sven Kulikowsky, CIO, und Michaela Brehm, Senior Manager Retail Solutions & Development, ANWR GROUP.

Was war das konkrete Problem, bevor ihr nach einer Lösung gesucht habt?

> „Die ANWR besteht aus vielen Verbundgesellschaften, und für unsere Händler war es schwer zu verstehen, wo sie welchen Service herbekommen. Wir wollten wissen: Wie können wir unseren Kunden unseren Service einfacher zur Verfügung stellen."

Sven Kulikowsky, CIO, ANWR GROUP

Wann wusstet ihr, dass octonomy wirklich einen Unterschied macht?

> „Das war in dem Moment, als wir merkten, dass unsere Lösung tatsächlich korrekt auf die Fragen antwortet, die der Händler stellt, aber gewisse Fragen außerhalb unseres Bereichs eben nicht beantwortet. Das war der Moment, als klar wurde: Das funktioniert."

Sven Kulikowsky, CIO, ANWR GROUP

Was würdet ihr Kolleginnen und Kollegen raten, die vor einer ähnlichen Entscheidung stehen?

> „Einfach machen, starten, probieren. Mit einem begrenzten POC, einem konkreten Thema anfangen und darauf sukzessive aufbauen."

Michaela Brehm, Senior Manager Retail Solutions & Development, ANWR GROUP

Wissen, das nur an einzelnen Ansprechpartnern hängt, ist kein System. Es ist ein Risiko.

celebrate company

celebrate COMPANY

~11 %
des Kontaktvolumens automatisiert

70 %
Automatisierungsquote in diesen Use Cases

98 %
korrekte Entscheidungen des Systems



„octonomy ist eine Plattform, die uns hilft, Kundenanfragen schnellstmöglich und maximal automatisiert mithilfe von KI zu beantworten."

Benedikt Peine, AI Officer, celebrate company

AUF EINEN BLICK

Das Unternehmen	Unternehmensgruppe für hochpersonalisierte Papeterie und Fotoprodukte mit den Marken kartenmacherei, faireparterie, Rosemood, celebrate Apps und mintkind. Eingesetzt im Kundenservice über Chat, E-Mail und Telefon.
Ausgangslage	Das monatliche Kontaktvolumen der kartenmacherei erhöht sich in der Peak Season um rund 200 %. Bislang wurde dieser Peak vor allem über zusätzliche Saisonkräfte abgefedert. Zugleich ist die Fehlertoleranz gering: Bei hochpersonalisierten, emotionalen Produkten kann sich das Unternehmen keine KI-Antworten leisten, die den hohen NPS-Wert gefährden.
Was celebrate company gemacht hat	celebrate company hat gemeinsam mit octonomy die Kontaktgründe analysiert und sich bewusst auf zwei Use Cases konzentriert: Sendungsverfolgung („Wo ist mein Paket?") und allgemeine FAQ-Anfragen. Im November 2025, mitten in der Weihnachtssaison, ging zunächst ein Chatbot live, kurz darauf ein E-Mail-Bot. Der Rollout erfolgte iterativ, beginnend mit einer Marke in einem kleineren Markt.
Ergebnis	70 % Automatisierungsquote bei Sendungsverfolgung und FAQ, dabei 98 % korrekte Entscheidungen. Die freigewordene Kapazität ist in Wissensmanagement, Dokumentation und Prozessoptimierung reinvestiert.

VIDEOS UND MATERIAL

SUCCESS STORY

MASTERCLASS

Die Case Study

AUSGANGSLAGE

celebrate company vereint mehrere Marken: die kartenmacherei als bekannteste Marke in Deutschland, ihren französischen Ableger faireparterie, außerdem celebrate Apps, mintkind und Rosemood. Im Zentrum steht hochpersonalisierte Papeterie, Hochzeitseinladungen, Geburtskarten, Weihnachtskarten, sowie Fotoprodukte wie Fotokalender und Fotobücher. celebrate Apps ergänzt das Portfolio um datenschutzkonformes Fotosharing für Familien und Events, mintkind um liebevoll gestaltete Erinnerungsprodukte und Geschenkartikel wie Babytagebücher und Freundebücher. Das Unternehmen produziert an zwei eigenen Standorten in Deutschland und Frankreich und begleitet jährlich über eine Million Menschen bei diesen Anlässen. Der NPS liegt bei 83, der B2C-E-Commerce-Benchmark bei 30 bis 45. Diesen NPS-Wert zu halten, während Standardanfragen zunehmend automatisiert werden, war für celebrate company die zentrale Anforderung an jede KI-Lösung.

HERAUSFORDERUNG

Bei der kartenmacherei verdreifacht sich das monatliche Kontaktvolumen im Dezember. celebrate company hat diesen Peak in der Vergangenheit vor allem über zusätzliche Saisonkräfte abgefedert, die rekrutiert, eingearbeitet und betreut werden mussten, ein erheblicher Aufwand für Service-Unit, People-Team und IT. Das Unternehmen hatte zuvor bereits mehrfach nach einem Anbieter gesucht, der das Kontaktvolumen automatisiert reduziert, aber lange keinen gefunden, dessen Antwortgenauigkeit dem eigenen Anspruch entsprach.

ZIELE

celebrate company hat vor dem Rollout klar unterschieden, wo KI eingesetzt werden soll und wo der Mensch im Kontakt bleibt:

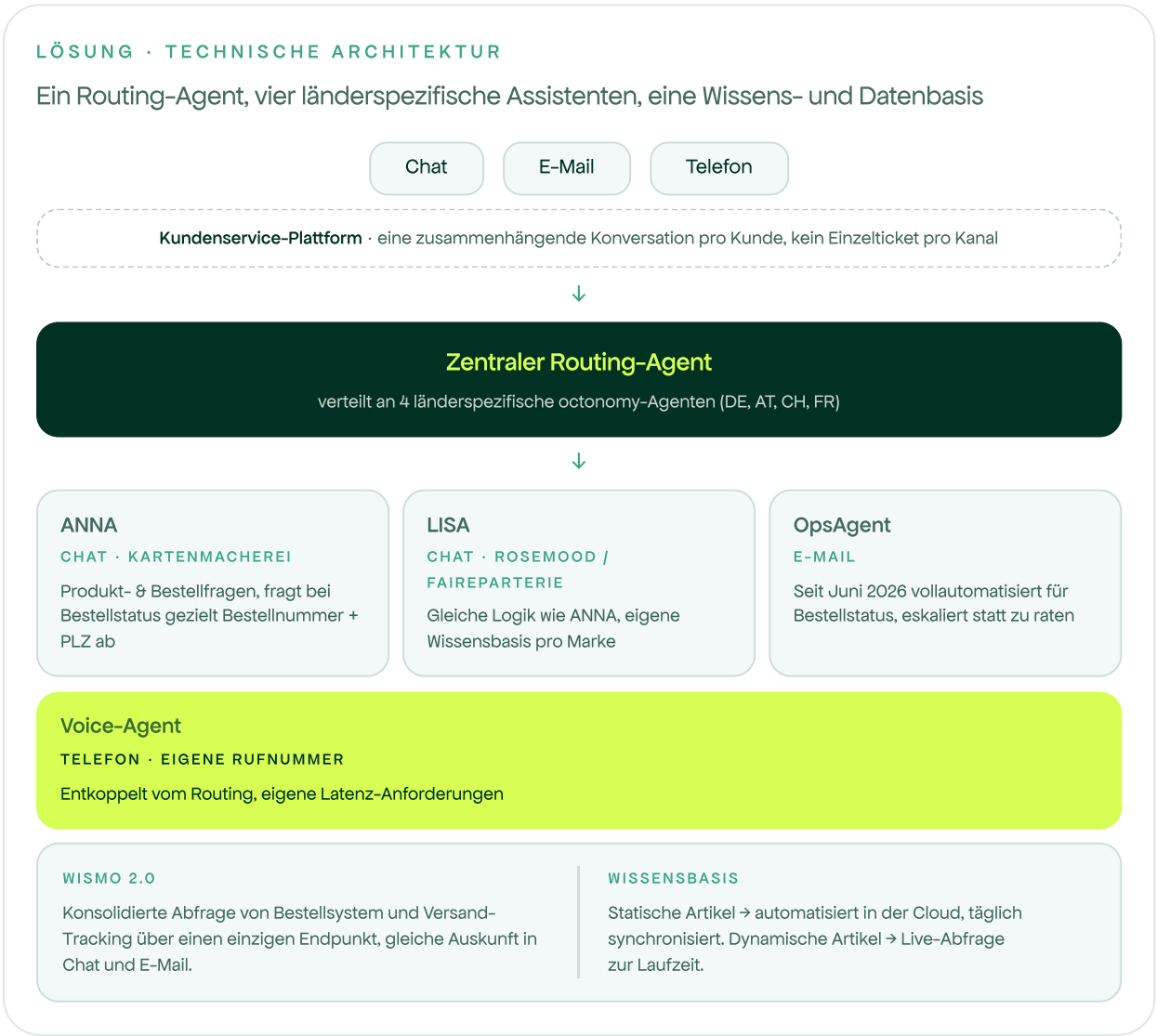
- **Standardanfragen automatisieren.** Sendungsverfolgung, der eCommerce-Klassiker „Wo ist mein Paket?“, sowie allgemeine FAQ-Anfragen wie Lieferzeiten oder Produktoptionen.
- **Antwortgenauigkeit priorisieren.** Eine hohe Korrektheit der Antworten war der ausschlaggebende Punkt für die Zusammenarbeit mit octonomy.
- **Den Menschen bei kritischen Fällen behalten.** Wenn zwei Tage vor einem Event eine Karte nicht angekommen ist oder in der Druckerei etwas kaputtgegangen ist, soll weiterhin ein Mensch reagieren.

LÖSUNG

celebrate company hat sich auf zwei Kontaktgründe konzentriert: Sendungsverfolgung und allgemeine FAQ-Anfragen. Zusammen machen diese beiden Use Cases rund 16 % des gesamten Kontaktvolumens über alle Kanäle, Chat, E-Mail und Telefonie, aus.

Umgesetzt wurde zunächst ein Chatbot, kurz darauf ein E-Mail-Bot. celebrate company wollte dabei ausdrücklich keine unterschiedlichen Lösungen für unterschiedliche Kanäle, sondern einen Anbieter, der alle Kanäle über eine Plattform bedient.

In den beiden Use Cases Sendungsverfolgung und FAQ automatisiert celebrate company rund 70 % der Anfragen. Bei allen Anfragen in diesen Use Cases traf das System in 98 % der Fälle die richtige Entscheidung: Es hat entweder korrekt geantwortet oder ein Anliegen, das es nicht kannte oder für das es nicht trainiert war, an einen Menschen weitergeleitet.



UMSETZUNG

Die Zusammenarbeit mit octonomy begann im Sommer 2025 mit einer gemeinsamen Prozessanalyse: Welche Kontaktgründe gibt es, welche eignen sich für Automatisierung, wo muss der Mensch im Loop bleiben.

Der Rollout erfolgte iterativ: celebrate company startete mit einer Marke in einem kleineren Markt und weitete den Einsatz danach schrittweise aus. Als der Chatbot funktionsfähig war, wollte das Team direkt die E-Mail-Automatisierung nachziehen, musste die Basis dafür aber zunächst noch anreichern, um von Anfang an die gewünschte Antwortqualität zu erreichen.
Der Go-Live erfolgte im November 2025, mitten in der Weihnachtssaison – ein Zeitpunkt, der intern als riskant galt, da sich das Kontaktvolumen in diesem Zeitraum verdreifacht. Er gelang dennoch innerhalb der laufenden Saison.

Bei celebrate company kümmern sich ein bis zwei Personen fast ausschließlich um das System, das intern als eigenständiges Agententeam verstanden wird. Je nach Projektphase stimmt sich dieses Team wöchentlich bis zweiwöchentlich mit octonomy ab.

UMSETZUNG · ZEITLEISTE

Von der ersten Prozessanalyse bis zur vollautomatisierten E-Mail-Bearbeitung

SOMMER 2025
Projektstart
Gemeinsame Prozessanalyse mit octonomy, Prozesse werden vor dem KI-Einsatz neu strukturiert

→

NOVEMBER 2025
Go-Live Chatbot
Mitten in der Weihnachtssaison, kurz darauf folgt der E-Mail-Bot

→

JUNI 2026
OpsAgent vollautomatisiert
E-Mail-Bestellstatus wird vollautomatisiert beantwortet, mit gezielter Eskalation an Menschen in Ausnahmefällen

ERGEBNIS

In den beiden Use Cases Sendungsverfolgung und FAQ automatisiert celebrate company rund 70 % der Anfragen. Bei allen Anfragen in diesen Use Cases trifft das System in 98 % der Fälle die richtige Entscheidung: Es beantwortet die Anfrage korrekt oder leitet sie an einen Menschen weiter.

celebrate company investiert die durch die Automatisierung freigewordene Zeit – automatisiert sind rund 11 % des gesamten Kontaktvolumens – in Wissensmanagement, Dokumentation für die KI und die laufende Prozessoptimierung. Für Benedikt Peine ist das kein abgeschlossenes Projekt, sondern eine dauerhafte Aufgabe: die Antwortqualität der KI regelmäßig überprüfen und das System weiter anreichern.

AUSBLICK

Bei den bisher rund 11 % automatisiertem Kontaktvolumen will celebrate company nicht aufhören. Die nächsten Schritte sind konkret:

- Bestehende Use Cases weiter optimieren, insbesondere die Antwortgenauigkeit kontinuierlich erhöhen
- Telefonie auf weitere Marken ausbauen, aktuell für kartenmacherei in DE, AT und CH im Einsatz, der Rollout für faireparterie und Rosemood läuft
- Erweiterte Use Cases wie Bestelländerungen und Zahlungsinformationen

Dabei hatte sich celebrate company als Ziel gesetzt, rechtzeitig vor der nächsten Weihnachtssaison so weit zu sein, dass zusätzliche Saisonkräfte in diesem Umfang nicht mehr erforderlich sind.

Our Side — Wie wir es gebaut haben

Ausgangslage: vier Ländermärkte, eine Frage dominiert den Support

celebrate company betreibt personalisierte Druckprodukte im E-Commerce. Im Kundenservice-Scope dieses Projekts liegen drei Marken, kartenmacherei, Rosemood und faireparterie, mit Präsenz in Deutschland, Österreich, der Schweiz und Frankreich. Von Anfang an war klar, dass eine einzelne Frage den Support dominiert, die Frage nach dem Verbleib der Bestellung, kurz WISMO. Diese Anfragen erforderten bislang eine manuelle Abfrage in mehreren Systemen, im Bestellsystem und beim Versand-Tracking-Anbieter, und machten einen erheblichen Teil des gesamten Supportaufkommens aus. Verschärft wird das durch eine ausgeprägte Saisonalität. Das Weihnachtsgeschäft von Oktober bis Dezember ist die Hochsaison, und celebrate company muss seine Personalplanung dafür bereits im Sommer festlegen.

Architekturprinzip: ein Routing-Agent für vier Ländermärkte

celebrate company nutzt eine Kundenservice-Plattform als zentrale Grundlage. Sie bildet sämtliche Chat-, E-Mail- und Telefonkontakte eines Kunden als eine zusammenhängende Konversation ab, während ein klassisches Ticket-Tool daraus Einzeltickets machen würde. Statt für jede Marke und jedes Land einen eigenen, einzeln lizenzierten Agenten auf dieser Plattform zu betreiben, haben wir uns für einen zentralen Routing-Agenten entschieden, der eingehende Konversationen an vier länderspezifische octonomy-Agenten weiterleitet. Für den Voice-Agenten haben wir bewusst ein eigenständiges, von diesem Routing entkoppeltes Setup mit eigener Rufnummer gebaut, da Telefonie andere technische Anforderungen an Latenz und Gesprächsführung stellt als Chat oder E-Mail.

Wissensbasis: das Support-Artikel-Tool als Grundlage, mit einer wichtigen Weichenstellung

Als Wissensquelle nutzen wir das bestehende Support-Artikel-Tool von celebrate company. Dabei unterscheiden wir zwei Arten von Inhalten. Statische Artikel ändern sich selten, weshalb wir sie automatisiert extrahieren und in einer Cloud-Wissensbasis ablegen, die täglich synchronisiert wird. Dynamische Artikel ändern sich häufiger, deshalb fragen wir sie zur Laufzeit direkt per Stichwortsuche aus dem Support-Artikel-Tool ab. In der Praxis lieferte die dynamische, stichwortbasierte Suche schlechtere Ergebnisse als die semantische Suche in der statischen Wissensbasis, weshalb wir den Großteil der Artikel in die statische, semantisch durchsuchbare Wissensbasis überführt haben.

Die Bausteine im Detail

BAUSTEIN	ZIELGRUPPE	AUFGABE
ANNA / LISA	Endkunden (Chat)	Beantworten Produkt- und Bestellfragen aus der Wissensbasis, leiten bei Bedarf an einen Menschen weiter
OpsAgent (E-Mail)	Endkunden (E-Mail)	Bearbeitet E-Mail-Anfragen über die Plattform, seit Juni 2026 vollautomatisiert für Bestellstatus
WISMO 2.0	Chat- und E-Mail-Agent	Konsolidierte Abfrage von Bestell- und Versanddaten für beide Kanäle
Voice-Agent (Telefon)	Endkunden	Telefonischer Support für die kartenmacherei-Hotline in DE, AT und CH, Bestellstatus und FAQ

ANNA und LISA sind die Chat-Assistenten für kartenmacherei beziehungsweise für Rosemood und fai-reparterie. Sie beantworten Produkt- und Bestellfragen aus der Wissensbasis, fragen bei Bestellstatus-anfragen gezielt nach Bestellnummer und Postleitzahl zur Verifikation und leiten bei Bedarf an einen Menschen weiter.

WANN DER AGENT AN DAS SERVICETEAM ABGIBT

01

Angaben unklar

Bestellnummer oder Postleitzahl nicht eindeutig, gezielte Rückfrage statt Auskunft auf Verdacht

02

Anliegen fremd

Kontaktgrund außerhalb von Sendungsverfolgung und FAQ, Weiterleitung mit Gesprächsverlauf

03

Nicht freigegeben

Fall bekannt, aber ohne Trainingsstand, Übergabe an den Menschen und Aufnahme in die Wissensbasis

Der OpsAgent übernimmt die E-Mail-Seite und ist seit Juni 2026 für Bestellstatusanfragen vollständig automatisiert. Im laufenden Betrieb sind uns einige Feinheiten aufgefallen, die sich erst im echten Kundenkontakt zeigen und die wir entsprechend nachgeschärft haben. Geht Geschlecht oder Namensform aus einer E-Mail nicht eindeutig hervor, greift jetzt eine verlässliche, neutrale Anrede. Liegt ein Sendungsstatus einmal nicht zeitgerecht vor, eskaliert der Assistent gezielt an einen Menschen, statt zu raten. Tracking-Links fügt er inzwischen einheitlich ein, die Ausgabe kommt konsequent ohne Emojis, wie es dem Ton der Marke entspricht, und reine Bestellbestätigungsanfragen gibt er direkt an einen Menschen weiter. Jede vollautomatisierte Antwort trägt außerdem eine länderspezifische Signatur mit Hinweis auf die KI-Generierung. Das entspricht den Transparenzpflichten aus Artikel 50 der EU-KI-Verordnung, wonach Kundinnen und Kunden erkennen müssen, dass sie mit einem KI-System kommunizieren.

WISMO 2.0 ist die gemeinsame technische Grundlage hinter den Bestellstatusauskünften in Chat und E-Mail. Es bündelt Bestellsystem und Versand-Tracking in einer Abfrage über einen einzigen Endpunkt. So geben Chat und E-Mail dieselbe Auskunft, auch wenn ein Kunde denselben Vorgang in beiden Kanälen anspricht.

Der Voice-Agent beantwortet Bestellstatus- und FAQ-Anfragen für die kartenmacherei-Hotline in Deutschland, Österreich und der Schweiz, der Rollout für faireparterie und Rosemood läuft. Telefonie verträgt kaum Wartezeit, denn ein Anrufer erwartet eine Antwort in Sekunden, während für einen Chat ein langsameres Tempo noch akzeptabel wäre. Deshalb haben wir den Assistenten mit schlankeren, gezielteren API-Abfragen ausgestattet als Chat und E-Mail. Die verbleibende Antwortzeit von 7 bis 10 Sekunden betten wir in natürlich wirkende Zwischenansagen ein, damit sich das Gespräch wie eine kurze Denkpause anfühlt und nicht wie eine technische Verzögerung.

Zwei Kennzahlen, die man nicht verwechseln sollte

celebrate company blickt auf zwei unterschiedliche Erfolgsmaße, und die Unterscheidung hilft, die Zahlen richtig einzuordnen. Das erste Maß ist die Leistung eines einzelnen Chatbots für sich, also wie viele Konversationen er eigenständig löst, ohne Weiterleitung an einen Menschen. Direkt nach der Wissensbasis-Migration von ANNA Deutschland lag dieser Wert bei rund 90 Prozent, ein Ergebnis auf Kanal-Ebene.

Das zweite Maß betrachtet das gesamte Kontaktvolumen. Die beiden automatisierten Use Cases Sendungsverfolgung und FAQ machen zusammen rund 16 Prozent aller Kontakte über Chat, E-Mail und Telefon aus. Innerhalb dieser beiden Use Cases automatisiert das System rund 70 Prozent der Anfragen, was etwa 11 Prozent des gesamten Kontaktvolumens entspricht. Bei allen Anfragen in diesen Use Cases trifft das System in 98 Prozent der Fälle die richtige Entscheidung: Es antwortet korrekt oder leitet an einen Menschen weiter.

Customer Side — Interview mit celebrate company

Auszüge aus den Gesprächen mit Benedikt Peine, AI Officer, und Michael Lindemeier, Unit Lead Service, celebrate company.

Was hat euch bei der Entscheidung für octonomy überzeugt?

> „octonomy ist eine Plattform, die uns hilft, Kundenanfragen schnellstmöglich und maximal automatisiert mithilfe von KI zu beantworten."

Benedikt Peine, AI Officer, celebrate company

Wie hat euer Team auf die Einführung reagiert?

> „Als die Kollegen gesehen haben, wie gut das Ganze nach dem Go-Live antwortet, war das eine positive Überraschung. Eine Kollegin hat sogar gesagt: octonomy, der neue Mitarbeiter des Monats."

Benedikt Peine, AI Officer, celebrate company

Woran messt ihr, ob es funktioniert?

> „Wir haben ganze 70 Prozent aller Anfragen im Bereich »Wo ist mein Paket?« und FAQs mit Hilfe von octonomy automatisieren können. Und von diesen 70 Prozent hat die KI in 98 Prozent der Fälle die richtige Entscheidung getroffen."

Benedikt Peine, AI Officer, celebrate company

Was verändert sich für das Serviceteam?

> „Mit octonomy automatisieren wir genau die Anfragen, die sich täglich wiederholen. Das gibt meinem Team den Kopf frei für die Fälle, bei denen es wirklich zählt: wenn zwei Tage vor der Hochzeit die Einladungen fehlen. Genau da entsteht der Service, für den unsere Kund:innen uns weiterempfehlen."

Michael Lindemeier, Unit Lead Service, celebrate company

Ein Rollout in der Hochsaison ist ein Stresstest. celebrate company hat ihn genutzt, nicht vermieden.

Manufacturing

Technisches Wissen in der Fertigung steckt oft in Schaltplänen, CAD-Dateien und annotierten Zeichnungen, also in Formaten, die generische KI-Lösungen nicht lesen, weil sie nur Text verarbeiten. Eine einzelne Antwort braucht häufig mehrere Quellen gleichzeitig, das passende Servicehandbuch, die Fehlerhistorie im ERP und den Wartungsstand im CRM. Ein Teil dieses Wissens ist zudem personengebunden. Erfahrene Servicetechniker gehen in Rente, und ihr Wissen geht mit ihnen, sofern es nicht vorher gesichert wird. Gleichzeitig kostet jede Minute Stillstand Geld und erhöht den Druck auf ohnehin knappe Serviceteams. octonomy liest technische Dokumentation als Bild statt nur als Text und liefert geprüfte, quellenbasierte Antworten.

04 & 05 · BORA und ALUCA zeigen, wie sich technischer Support darauf vorbereiten lässt.

BORA

BORA

4x	NPS 80	>3 Mon.
mehr Chatbot-Traffic	gehalten	Einarbeitung vorher



„Wir haben das Potenzial, Innovationen zu bringen, aber wir müssen uns auch mal trauen."

Florian Plenk, Head of Technical Support, BORA

AUF EINEN BLICK

Das Unternehmen	Deutsch-österreichischer Hersteller von Premium-Küchengeräten, bekannt für Kochfelder mit integrierter Ablufttechnik, tätig in 58 Ländern. Eingesetzt im technischen Kundendienst und im Webchat.
Ausgangslage	Bei linearer Entwicklung hätte das starke Wachstum einen massiven Ausbau der Servicekapazität erfordert, und das in einer Region mit knappen Fachkräften. Das Onboarding neuer Customer-Support-Mitarbeitender dauert mehr als 3 Monate.
Was BORA gemacht hat	Schrittweiser Aufbau von drei Einsatzbereichen, einem internen KI-Assistenten für den Customer Support, intern Gregario genannt, einem Chatbot für die Homepage und einem Agenten für externe Servicetechniker, der bereits im Einsatz ist.
Ergebnis	Viermal mehr Chatbot-Traffic auf der Homepage, geringerer Kapazitätsbedarf im Customer Support, schnelleres Onboarding, ein gehaltener NPS von 80 und sofortige Mehrsprachigkeit.

VIDEOS UND MATERIAL

CASE STUDY

MASTERCLASS

SUCCESS STORY

Die Case Study

AUSGANGSLAGE

BORA ist ein deutsch-österreichischer Hersteller hochwertiger Küchengeräte, insbesondere Kochfelder mit integrierter Ablufttechnik und Dampfbacköfen, gegründet von Willi Bruckbauer. BORA gilt heute als Gattungsbegriff für diese Technologie, mit Sitz im Inntal nahe Kufstein und Rosenheim. Das Unternehmen beschäftigt aktuell rund 850 Mitarbeitende und will in diesem Jahr auf rund 1.000 wachsen. Service und Kundendienst umfassen rund 120 Personen, inklusive Customer Support und eigener Servicetechniker.

BORA wächst stark und plant, die Mitarbeiterzahl aufgrund neuer Produkte und Märkte deutlich zu erhöhen. Für den technischen Kundendienst hätte das bei linearem Wachstum eine Verdoppelung bis Vervierfachung der Servicekapazität bedeutet, und das in einer Region, in der Fachkräfte schwer zu finden sind, insbesondere mehrsprachige. Das Onboarding neuer Customer-Support-Mitarbeitender dauerte mehr als 3 Monate, weil das Wissen verteilt und schwer zugänglich war.

HERAUSFORDERUNG

Florian Plenk leitet den technischen Service bei BORA seit sechs Jahren. Als er sich vor rund zwei Jahren angeschaut hat, was das geplante Wachstum für seinen Bereich bedeutet, war die Rechnung ernüchternd, zwischen Faktor zwei und vier mehr Serviceeinsätze.

> „BORA ist extrem dynamisch und extrem ambitioniert, wir wollen sehr stark wachsen. Und ich habe mir mal angeschaut, wie viele Serviceeinsätze da auf uns zukommen. Da kam irgendwas zwischen Faktor 2 und 4 raus. Also war der Gedanke klar: Wir müssen irgendwie automatisieren, weil linear wachsen geht so nicht."
>
> Florian Plenk, Head of Technical Support, BORA

BORA hat aktuell 24 Produkte, zu jedem mehr als 100 Seiten Dokumentation. Neue Customer-Support-Mitarbeitende brauchen mehr als 3 Monate, um wirklich fit zu sein. Gleichzeitig wächst BORA international. Servicepartner in Frankreich, Italien oder Skandinavien betreuen neben BORA fünf bis sechs weitere Marken. Englisch funktioniert dabei oft nicht, und das Wissen war ausschließlich auf Deutsch dokumentiert.

ZIELE

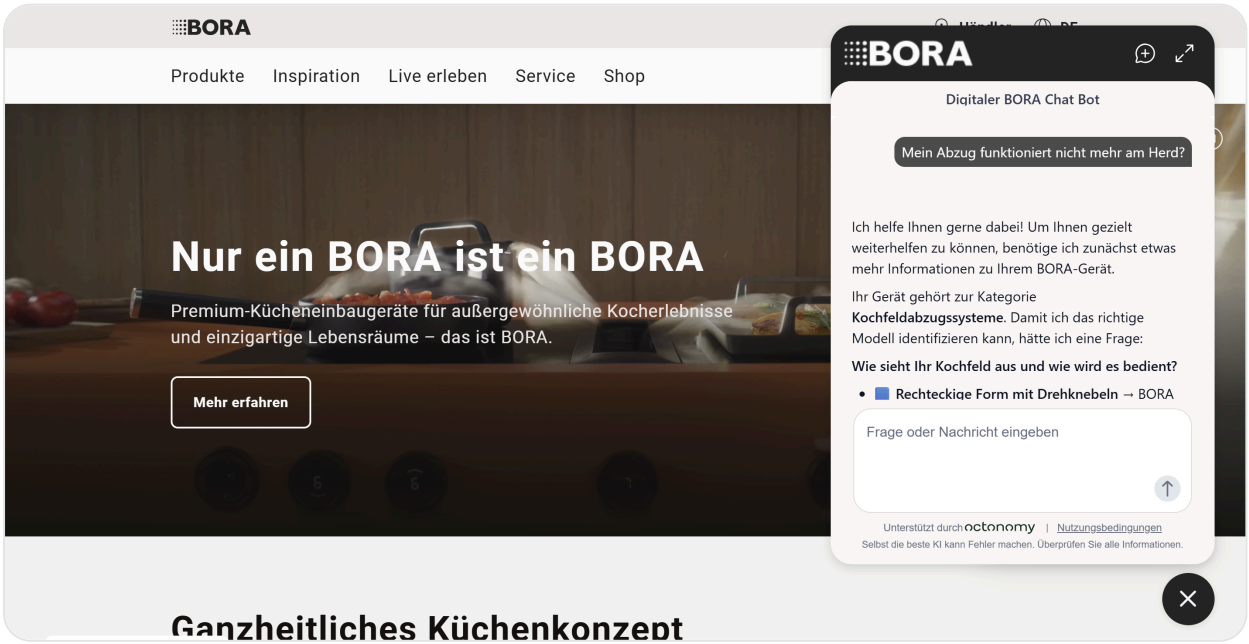
BORA hat drei klare Prioritäten gesetzt. Die erste ist ein schnelleres Onboarding, bei dem neue Customer-Support-Mitarbeitende über besseren Zugang zu Wissen schneller produktiv werden, ohne dass die Komplexität sinkt. Die zweite ist das Abschirmen von Standardanfragen, denn Fragen wie „Wie reinige ich mein Kochfeld“ kommen täglich zigfach, und kein Mensch sollte sie noch beantworten müssen. Die dritte ist die Internationalisierung, damit Servicepartner und Techniker in ganz Europa in ihrer eigenen Sprache auf dasselbe Wissen zugreifen, ohne dass BORA dafür mehrsprachige Spezialkräfte einstellen muss.

LÖSUNG

BORA nennt seinen internen KI-Assistenten Gregario, nach dem Wasserträger im Radsport, der immer da ist, immer schnell und immer alles findet. Gregario ist im Customer Support eingebettet und beantwortet Wissensfragen zu allen Produkten und Servicethemen, auf Basis von Dokumenten aus SharePoint, mittlerweile angebunden an Pimcore als Single Source of Truth. Insbesondere für neue Mitarbeitende ist Gregario zum täglichen Werkzeug geworden, während erfahrene Kolleginnen und Kollegen ihn seltener nutzen. Genau das war der Plan, denn Gregario übernimmt die Routine, damit sich die erfahrenen Kräfte um das kümmern können, was wirklich Kompetenz erfordert.

Nach dem internen Rollout im Oktober und November hat BORA im Januar denselben Wissensstand als Chatbot auf die Homepage gebracht, ohne interne Informationen und mit angepassten Leitlinien für den Kundenkontakt. Inzwischen ist auch ein Chatbot für externe Servicetechniker und Servicepartner im Einsatz. Ein französischer Techniker gibt auf Französisch ein, wo das Problem liegt, und erhält eine verlässliche Anleitung in seiner Sprache.

BEISPIEL AUS DEM BETRIEB



Screenshot: bora.com

UMSETZUNG

BORA hat bewusst mit dem internen Chatbot begonnen, um in einem kontrollierten Umfeld zu validieren, denn im Customer Support sitzen Leute, die wissen, ob eine Antwort stimmt. Von der Entscheidung bis zum internen Go-Live im Oktober und November vergingen rund dreieinhalb Monate. Der ursprüngliche Plan war, PDFs einfach zu importieren, doch das erwies sich als zu einfach, weil sich PDFs nicht automatisch aktuell halten lassen. Die Entscheidung, Inhalte über Pimcore zu kanalisieren, war richtig, kostete aber zusätzliche Zeit, weil sie im ursprünglichen Plan nicht vorgesehen war.

> „Ursprünglich hat man mal gemeint, okay, wir packen hier 20, 30 PDFs rein und dann läuft die Sache. Es hat sich herausgestellt, ganz so einfach ist es nicht. Was aber auch wieder für octonomy spricht, weil sonst könnte man es auch selbst machen."
>
> Florian Plenk, Head of Technical Support, BORA

Für Florian Plenk ist der Go-Live ein Startschuss und kein Schlusspunkt. BORA nimmt heute Senior-Mitarbeitende aus dem Tagesgeschäft heraus, um sie an Gregario weiterarbeiten zu lassen, und etwa die Hälfte der eingesparten Kapazität fließt so wieder ins System zurück. Bei der Entscheidung für octonomy gaben zwei Faktoren den Ausschlag. Der erste war die ehrliche Kommunikation darüber, was die Plattform bereits leisten konnte und was nicht. Der zweite war die Breite der Plattform, mit der BORA heute eigene Agenten über die grafische Oberfläche baut. Auch Change Management war zentral, denn BORA konnte glaubwürdig kommunizieren, dass es keinen Stellenabbau gibt, sondern weniger Neueinstellungen.

ERGEBNIS

Der Homepage-Chatbot verzeichnet viermal mehr Traffic als der vorherige, starre Baumstruktur-Chatbot. Im Customer Support sinkt der Kapazitätsbedarf spürbar, trotz laufendem Betrieb. Das Onboarding neuer Mitarbeitender ist spürbar beschleunigt. Der NPS liegt bei 80 und wurde trotz Automatisierung gehalten und im laufenden Jahr leicht gesteigert. Mehrsprachigkeit ist sofort verfügbar. Der Servicetechniker-Chatbot in Landessprache ist bereits im Einsatz. Die nächsten Schritte sind konkret, die vollautomatische Beantwortung von Standardanfragen, die automatische Ticketvalidierung und Rechnungsprüfung sowie ein KI-basiertes Briefing für Servicepartner vor jedem Einsatz.

Our Side — Wie wir es gebaut haben

BORA ist ein Hersteller von Premium-Küchengeräten mit Sitz in Österreich.

Der Kundenservice ist in einen First und Second Level gegliedert. Rund 40 Mitarbeitende arbeiten im First Level und betreuen die Endkunden direkt – bei einer Kundenzufriedenheit von 4,8 von 5. Insgesamt ist der Servicebereich mit dem Wachstum des Geschäfts in sechs Jahren von 35 auf 120 Mitarbeitende gewachsen. Damit stellt sich die Frage: Soll die Serviceorganisation auch künftig linear mit dem Geschäft mitwachsen?

Ausgangslage: drei Quellen, drei Logiken, eine Frage

Die Anfragen sind ungewöhnlich anspruchsvoll für ein Assistenzsystem. Es geht selten um einen Bestellstatus und häufig um ein Gerät, dessen Modellbezeichnung der Kunde nicht kennt, in einer Fehlersituation, die er nur beschreiben kann.

Das nötige Wissen verteilt sich auf drei Quellen. Das Produktwissen liegt im zentralen Produktinformationssystem, in dem BORA Handbücher, Datenblätter und Serviceanleitungen pflegt. Die Vorgänge liegen im CRM-System. Die Diagnosewege liegen in einem eigens gepflegten Fehlerlösungsbaum.

Architekturprinzip: Ebenen statt Kanäle

Wir haben BORAs Assistenten nicht nach Kanälen geschnitten, sondern nach dem Vorwissen der Person, die fragt. Dieselbe Wissensbasis, drei Zuschnitte.



Auf Ebene 0 spricht der Endkunde. Der Assistent führt ihn über einen strukturierten Entscheidungsbaum mit Produktbildern zunächst zur richtigen Gerätegruppe und antwortet erst danach inhaltlich, weil eine Antwort zum falschen Gerät weniger hilft als eine gezielte Rückfrage.

Auf Ebene 1 spricht die Servicemitarbeiterin. Sie hat den Kunden am Telefon und braucht in Sekunden die richtige Anleitung.

Auf Ebene 2 spricht der Servicetechniker vor Ort. Für ihn dienen dieselben Dokumente als Reparaturanweisung.

Den Agenten für die dritte Ebene hat BORA selbst aufgesetzt. Der zuständige Kollege kopierte dafür den Agenten der Ebene 1 und passte dessen Grundanweisung an. Dass dieser Schritt ohne octonomy möglich war, entspricht dem Ziel des Projekts.

Bei der Mehrsprachigkeit haben wir bewusst auf einen Assistenten pro Sprache und auf einen vorgeschalteten Verteiler verzichtet. Stattdessen gibt es einen deutschsprachigen Assistenten, der bei Bedarf übersetzt. So bleibt ein einziger Agent in Pflege, statt sieben, die mit der Zeit auseinanderlaufen.

ASSISTENT	ZIELGRUPPE	ZUGANG
Assistent Ebene 0	Endkunden	Öffentlicher Webchat, Deutsch/Englisch
Support Assistent Ebene 1	BORA Kundenservice	Intern, erweiterte Wissensbasis
Assistent Ebene 2	Servicetechniker im Außendienst	Im Einsatz, von BORA selbst aufgesetzt
TicketMaster	Endkunden	Abfrage von Ticketstatus & Lösungsfindung

Die Bausteine im Detail

Der Ticketmaster hilft Kunden bei einem bereits bestehenden Anliegen. Er prüft direkt über die Schnittstelle des CRM-Systems, ob zum Kundenanliegen schon ein Vorgang existiert. Der Zugriff ist identitätsgeprüft, sodass nur berechtigte Anfragen einen Vorgang auslösen. Den Link zum Vorgang erzeugt das CRM-System selbst, was eine zusätzliche Fehlerquelle spart.

Zusätzlich gibt es eine Suche auf der öffentlichen BORA-Website. Sie ist bewusst nachrangig eingestellt und greift erst, wenn die internen Quellen nichts liefern. Sie durchsucht die Seite live, ohne Inhalte zu speichern. Gemeinsam mit BORA haben wir festgelegt, welche Themen sie abdeckt: Veranstaltungen, Ratgeber und ergänzende Serviceinformationen. Die Produktdaten bleiben außen vor, weil sie in der Wissensbasis ohnehin besser gepflegt sind.

Der Assistent der Ebene 0 beantwortet Planungs- und Installationsfragen ausdrücklich nicht. Er verweist auf die Planungsfibel oder legt einen Vorgang an. Diese Grenze haben wir auf Wunsch von BORA gezogen, weil eine plausibel klingende, aber falsche Einbauempfehlung teurer ist als gar keine.

Das Fundament: die Arbeit, die man nicht sieht

Ein Assistent ist immer nur so gut wie die Information, die er findet. Ausschlaggebend dafür ist, wie sorgfältig das Dokument ausgezeichnet ist, in dem die Antwort steht. Gemeinsam mit BORA haben wir jedes Wissensdokument mit drei Angaben versehen: für welche Ebene es gilt, in welcher Sprache es vorliegt und zu welchem Produkt es gehört.

Parallel dazu haben wir die interne Objektkennung mit der Artikelnummer verbunden, über einen einmaligen Gesamtabgleich je Aktualisierungslauf, statt sie Dokument für Dokument nachzuschlagen. Die Artikelnummer ist die Nummer, mit der BORA und seine Kunden tatsächlich arbeiten. Die Aufbereitung der Wissensbasis hat anfangs mehrere Tage gedauert. Heute läuft sie in Stunden, nachts automatisch, mit einer festen wöchentlichen Aktualisierung. Zusätzlich haben wir den Fehlerlösungsbaum vollständig durchgesehen, unbefüllte Fehlerkennungen aufgedeckt und den Abgleich mit dem CRM-System bereinigt. Der Fehlerlösungsbaum ist per direkter Schnittstelle angebunden und liefert seine Informationen in Echtzeit.

Für den Kunden bleibt all das unsichtbar. Genau hier entscheidet sich jedoch, ob er eine richtige Antwort bekommt.

DIE WISSENSAUFBEREITUNG

01	02	03	04
Datenerfassung	**Auszeichnung**	**Zuordnung**	**Aktualisierung**
Technische Dokumentation, Fehlerlösungsbaum, CRM-Daten	Ebene, Sprache und Produkt je Dokument	Objektkennung je Artikelnummer	Nächtlicher Lauf, wöchentlicher Rhythmus

AUFBEREITUNGSDAUER mehrere Tage → wenige Stunden

Befähigung statt Abhängigkeit

Einen technischen Kollegen aus BORAs Serviceentwicklung haben wir so eingearbeitet, dass er die Plattform selbstständig bedient. Er passt Grundanweisungen an, prüft Wissensbasen und setzt, wie beschrieben, inzwischen eigene Agenten auf.

Was daraus entsteht

Der Wert dieses Projekts lässt sich nicht an der Zahl der Gespräche ablesen, auch wenn diese sich seit dem Frühjahr von rund 20 auf etwa 50 am Tag verdoppelt hat. Er entsteht an mehreren Stellen gleichzeitig:

- Die Servicemitarbeiterin findet die passende Anleitung, während der Kunde noch spricht.
- Der Techniker im Außendienst geht die Reparaturschritte schon auf dem Weg zum Einsatz durch.
- Der Kunde ordnet sein Gerät anhand eines Bildes zu, statt eine Modellbezeichnung suchen zu müssen.

Wie geht es weiter

Als nächster Schritt soll ein OpsAgent die E-Mail-Bearbeitung unterstützen. Geht eine Anfrage ein, erstellt dieser auf Basis derselben Wissensbasis einen Antwortentwurf. Eine Servicemitarbeiterin prüft und ergänzt ihn, bevor die Antwort versendet wird. Die Entscheidung bleibt damit zunächst bewusst beim Menschen. Ob der OpsAgent später selbstständig antworten kann, entscheidet BORA auf Grundlage der gesammelten Erfahrungen.

Customer Side — Interview mit BORA

Auszug aus dem Gespräch mit Florian Plenk, Head of Technical Support, BORA.

Welche zwei Aufgaben übernimmt octonomy bei euch zentral?

> „octonomy übernimmt zentral zwei Aufgaben. Das eine ist die Wissensausgabe, wir haben 24 Produkte und zu jedem Produkt etwa 100 Seiten Dokumentation. Das andere ist, auf der Homepage relativ simple Standardfragen zu beantworten."

Florian Plenk, Head of Technical Support, BORA

Wie groß ist der Kapazitätsdruck, über den wir hier sprechen?

> „Wir sind aktuell ungefähr 100 bis 120 Leute im Service, Customer Support und eigene Servicetechniker. Ich denke, wir würden durchaus mal das Doppelte brauchen."

Florian Plenk, Head of Technical Support, BORA

Was würdest du anderen raten, die vor einer ähnlichen Entscheidung stehen?

> „KI ist in aller Munde, aber das ist nicht der Heilsbringer in dem Sinn, dass ich sage, ich drücke jetzt aufs Knöpfchen und dann ist alles gut, sondern das ist ein Prozess, und dieser Prozess wird immer besser. Ich würde tatsächlich anfangen, einfach anfangen, lernen. Sonst ist man Nachzügler."

Florian Plenk, Head of Technical Support, BORA

Wachstum ohne mehr Personal ist kein Zufall. Es ist eine Entscheidung, die man vorbereiten muss.

ALUCA

20+	~28 %	4
europäische Länder	Kapazitätsbindung durch 2 FTE vorher	Rollout-Phasen



„Wir wollen ein Unternehmen sein, das KI nicht begleitet, sondern damit führt. Mit octonomy bauen wir die Grundlage dafür."

Frédéric Straß, CEO, ALUCA GmbH

AUF EINEN BLICK

Das Unternehmen	Anbieter hochwertiger Fahrzeugeinrichtungen aus Aluminium, seit 1993, mit Export in mehr als 20 europäische Länder. Eingesetzt im internationalen Support für Servicetechniker und Partner.
Ausgangslage	Kunden, Partner und Händler in mehr als 20 europäischen Ländern. Rund 35 Anrufe und 26 E-Mails pro Tag binden etwa 270 Minuten Kapazität, fast 28 Prozent der Arbeitszeit von zwei Mitarbeitenden im Produktmanagement.
Was ALUCA gemacht hat	Ergänzung des Supports um die octonomy KI-Expertenlösung, die Anfragen versteht, Informationen aus der Wissensdatenbank abruft und Prozesse im Idealfall end-to-end abschließt.
Ergebnis	Schnellere Bearbeitung von Standardanfragen, einheitlicher Wissensstand für Mitarbeitende, skalierbares Setup für weiteres internationales Wachstum.

VIDEOS UND MATERIAL

CASE STUDY

Die Case Study

AUSGANGSLAGE

Die ALUCA GmbH ist seit 1993 Anbieter hochwertiger Fahrzeugeinrichtungen aus Aluminium, spezialisiert auf die Bedürfnisse von Handwerkern und Technikern, Fuhrparkmanagern und Autohändlern. Hauptsitz und Produktionsstandort ist Rosengarten in Baden-Württemberg. Die Unternehmensgruppe beschäftigt über 200 Mitarbeitende und betreibt Vertriebsstationen in Berlin, Braunschweig und Olpe, dazu die Tochtergesellschaften Aluca Benelux B.V., Aluca UK LTD. und Aluca France SAS. Über den eigenen Vertrieb und ein breites Partnernetzwerk exportiert ALUCA in mehr als 20 europäische Länder.

ALUCA betreut Kunden, Partner und Händler in diesen mehr als 20 europäischen Ländern. Der Anspruch an Verfügbarkeit, Antwortqualität und Mehrsprachigkeit im Support wächst mit diesem Netzwerk. Das Wissen ist im Unternehmen vorhanden, aber nicht systematisch zugänglich.

HERAUSFORDERUNG

Der Anspruch der Kunden an zeitliche Verfügbarkeit, Antwortqualität und Mehrsprachigkeit im Kundenservice steigt stetig. Bei ALUCA läuft der Support unter hohem Druck. Rund 35 Anrufe und 26 E-Mails pro Tag binden etwa 270 Minuten Kapazität, was nahezu 28 Prozent der Arbeitszeit von zwei Mitarbeitenden im Produktmanagement entspricht. Gleichzeitig wächst das Partnernetz international. Um Qualität und Geschwindigkeit ohne proportionalen Personalaufbau zu sichern, setzt ALUCA auf die octonomy KI-Lösung.

ZIELE

Ziel des Projekts ist eine zentrale Wissensdatenbank, die durch KI-gestützte Systeme intern wie extern die Effizienz und Qualität des Supports steigert. Häufig gestellte Fragen sollen schnell und präzise beantwortet werden, komplexe Anfragen gezielt an die zuständigen ALUCA-Mitarbeitenden gehen, um eine optimale Betreuung sicherzustellen.



LÖSUNG

octonomy arbeitet als koordiniertes Netz aus spezialisierten KI-Agenten. Ein Supervisor-Agent kategorisiert jede eingehende Anfrage und leitet sie an den zuständigen Agenten weiter, für ALUCA sind das Agenten für Fehlerbehebung, Garantiefälle, Beschwerden, Terminplanung und Ersatzteile. Anfragen erreichen das System per Chat oder Webhook, von internen Mitarbeitenden ebenso wie von externen Partnern und Kunden. Die Wissensbasis kombiniert Vector Store und Graph-Datenbank, um Informationen kontextgenau abzurufen, und ist direkt mit den ALUCA-Systemen verbunden, darunter CRM, Ticketsystem und weitere Unternehmenssysteme.

UMSETZUNG

ALUCA und octonomy haben einen Rollout in vier aufeinander aufbauenden Phasen geplant. Phase eins legt das Fundament, indem Mitarbeitende Wissen per Text oder Spracheingabe erfassen und Spracheingaben automatisch transkribiert und in der Wissensdatenbank abgelegt werden. Phase zwei bringt die interne Einführung, bei der der KI-Agent für Mitarbeitende live geht und ein Feedback-System für kontinuierliche Verbesserung sorgt. Phase drei öffnet den Support nach außen, indem die Chatbot-Schnittstelle für Partner und ausgewählte Kunden live geht und unbeantwortete Fragen automatisch mit Kontext an den ALUCA-Kundensupport gehen. Phase vier erweitert das System um den ALUCA-eigenen Fahrzeugeinrichtungs-Konfigurator aluCLICK, also um individuelle Konfiguration und automatische Angebotserstellung direkt im KI-gestützten Support-Flow.

ERGEBNIS

Standardanfragen werden schneller bearbeitet, der Wissensstand ist einheitlich für Mitarbeitende, und das Setup ist skalierbar für das weitere internationale Wachstum. Mit der Integration von aluCLICK entsteht perspektivisch ein System, das Kundenprozesse von der Konfiguration bis zur Angebotserstellung vollständig automatisiert abschließt.

> „Wir wollen ein Unternehmen sein, das KI nicht begleitet, sondern damit führt. Mit octonomy bauen wir die Grundlage dafür."
>
> Frédéric Straß, CEO, ALUCA GmbH

Internationales Wachstum funktioniert nur, wenn das Wissen mitwächst.

Our Side — Wie wir es gebaut haben

Ausgangslage: Hilfe vor Ort, ohne digitale Diagnose

ALUCA baut Innenausstattungssysteme für Fahrzeuge und Wohnmobile, also Schubladen- und Regalsysteme, die Werkzeug und Material unterwegs griffbereit halten. Diese Systeme sind rein mechanisch. Es gibt keine eingebaute Diagnose, die einem Servicetechniker vor Ort sagt, warum zum Beispiel eine Schublade nicht in ein bestimmtes Regalsystem passt.

Bislang blieb in solchen Momenten nur ein Anruf beim ALUCA-Support, verbunden mit der Suche nach der einen Person im Unternehmen, die die Antwort kannte. Diese Person war nicht immer verfügbar, und es entstanden spürbare Wartezeiten. Genau hier setzen wir an, mit einer Anlaufstelle, die Servicetechnikern und Monteuren direkt vor Ort weiterhilft, ohne Warteschlange.

Unser Ansatz: erst Wissen schaffen, dann den Chatbot darauf aufbauen

Dieses Projekt unterscheidet sich von anderen in einem Punkt. ALUCA verfügte zu Beginn praktisch über keine digital nutzbare Wissensgrundlage, nur über einen PDF-Produktkatalog. Ein Chatbot allein auf dieser schmalen Basis hätte wenig gebracht. Wir haben das Projekt deshalb in zwei Teile gegliedert, die zusammen erst die eigentliche Lösung ergeben, ein Werkzeug, mit dem ALUCA selbst laufend neues Praxiswissen erfasst, und einen Chatbot, der auf diesem wachsenden Wissen aufsetzt.

BAUSTEIN	ZIELGRUPPE	AUFGABE
Wissenserfassungs-Frontend	ALUCA-Mitarbeitende	Erfassen neues Produkt- und Fehlerwissen per Sprachaufnahme oder strukturierter Eingabe
Chatbot	Servicetechniker, Monteure	Beantwortet Produkt- und Fehlerfragen auf Basis der freigegebenen Wissensbasis

Die Bausteine im Detail

Das Frontend haben wir gemeinsam mit unserem Partner Talentship entwickelt. Es ist über ALUCAs unternehmenseigenes Single-Sign-On angebunden, sodass sich Mitarbeitende mit ihrem gewohnten Firmenzugang anmelden. Darin gibt es zwei Rollen. Ein regulärer Nutzer kann neues Wissen auf zwei Wegen einbringen, als kurze Sprachaufnahme in einem Format, das an einen Podcast-Beitrag erinnert, oder als strukturierte Eingabe über ein Dropdown- und Kategoriesystem nach Produkt oder Fehlerbild.

Ein zweiter Nutzertyp ist der Approver. Er sichtet diese Einträge und gibt sie frei, bevor sie in der Wissensbasis landen. Diese Qualitätsschleife ist bewusst gesetzt, damit nicht jede Rohnotiz ungeprüft zur Wissensquelle wird. Erst freigegebene Einträge landen als Markdown-Datei in ALUCAs zentraler Dokumentenablage. Markdown ist ein einfaches, gut editierbares Textformat, das auch Bilder einbetten kann.

Von dort übernimmt unsere Plattform. Sobald eine freigegebene Markdown-Datei in der zentralen Dokumentenablage liegt, holt ein Trigger sie automatisch in die octonomy-Wissensbasis. Der Chatbot arbeitet ausschließlich auf dieser Basis und beantwortet Fragen von Servicetechnikern und Monteuren zu Produkten und typischen Fehlerbildern, direkt und ohne Warteschlange.

Wie wir mit den Daten umgegangen sind

Die vorhandenen Produktdaten liegen überwiegend als Tabellen in PDF-Form vor. Eine Anbindung an native, strukturierte Quellen wie Excel oder eine Datenbank wäre technisch die präzisere Variante gewesen, doch zu Projektstart fehlte dafür der Systemzugriff auf ALUCA-Seite. Wir haben uns deshalb für die PDFs entschieden und investieren gezielt mehr Aufwand in die Vorverarbeitung, damit die Tabelleninhalte saubere, verwertbare Antworten ergeben, mit einem Zielwert von 95 Prozent Antwortgenauigkeit.

Eine zweite, bewusste Entscheidung betrifft den Datenschutz. Ursprünglich hatten wir überlegt, Support-Telefonate mit Kunden aufzuzeichnen und zu transkribieren, eine ergiebige Wissensquelle. Nach Abstimmung mit ALUCAs Datenschutzbeauftragtem sind wir diesen Weg nicht weiter gegangen und haben mit dem Wissenserfassungs-Frontend eine datenschutzfreundlichere Alternative gebaut. Sie erreicht dasselbe Ziel, aktuelles Praxiswissen von den Menschen, die es täglich anwenden, direkt und strukturiert in die Wissensbasis zu bringen.

BRANCHE 04

Energie und Infrastruktur

Im technischen Außendienst bleibt Expertenwissen oft fragmentiert. Es steckt in Köpfen, PDFs oder E-Mails und ist im Einsatz selbst nicht verfügbar. Klassische KI-Ansätze sind bei komplexen oder sicherheitsrelevanten Fällen zu ungenau, und wenn erfahrene Fachkräfte zum Engpass werden, bleiben weniger erfahrene Kolleginnen und Kollegen unsicher. Gleichzeitig kostet jede ungeplante Stillstandszeit einer Anlage unmittelbar Geld. octonomy bündelt Wartungsanleitungen, Einsatzberichte und Erfahrungswissen zentral und macht sie im Einsatz nutzbar, ohne dass die bestehende Dokumentation neu geschrieben werden muss.

06 — Prokon zeigt, wie sich das auf 470 Windräder in 77 Windparks anwenden lässt.

Prokon

50 %	25 %	95 %	10 Tage
weniger Anfahrten	schnellere Problemlösung vor Ort	Genauigkeit bei komplexen Fragen	bis Machbarkeitsstudie



„Mit octonomy reduzieren wir unsere Anfahrten um 50 Prozent. Das System gibt uns die Sicherheit, die wir in der kritischen Infrastruktur brauchen."

Dr. Yousef Farschtschi, Bereichsleiter IT und Digitales, Prokon

AUF EINEN BLICK

Das Unternehmen	Deutschlands größte Energiegenossenschaft, betreibt Onshore-Windkraftanlagen in vier europäischen Ländern. Eingesetzt im Field Service zur Störungsdiagnose an Windkraftanlagen.
Ausgangslage	77 Windparks, 20 verschiedene Anlagengenerationen, 2.600 technische Dokumente mit über 50.000 Seiten. Durchschnittlich sechs Anfahrten pro Störung, weil Diagnosen vor Ort getroffen werden mussten und relevantes Wissen nicht greifbar war.
Was Prokon gemacht hat	Integration der octonomy KI-Plattform in Microsoft Dynamics Field Service. Das System analysiert Fehlercodes, verknüpft Monitoring-Daten mit Herstellerdokumentation und erstellt konkrete Handlungspläne, bevor Techniker zur Anlage fahren.
Ergebnis	Halbierte Anfahrten, 25 Prozent schnellere Problemlösung vor Ort, messbar höhere Anlagenverfügbarkeit.

VIDEOS UND MATERIAL

CASE STUDY

SUCCESS STORY

Die Case Study

AUSGANGSLAGE

Prokon erzeugt rund ein Prozent des deutschen Stroms und ist damit Deutschlands größte Energiegenossenschaft. In vier europäischen Ländern entwickelt, baut und betreibt das Unternehmen Onshore-Windkraftanlagen. Der erzeugte Strom wird an den Märkten verkauft und über ein eigenes B2C-Geschäft direkt an Verbraucherinnen und die über 50.000 Genossenschaftsmitglieder geliefert.

Prokon betreibt 77 Windparks mit 20 verschiedenen Anlagengenerationen und 2.600 technischen Dokumenten von über 50.000 Seiten. Fällt eine Anlage aus, zählt jede Minute. Techniker fahren oft mehrfach zur selben Anlage, weil die Diagnose vor Ort getroffen werden muss und das relevante Wissen nicht greifbar ist. Durchschnittlich sechs Anfahrten pro Störung bedeuten hohe Kosten, lange Stillstandzeiten und eine spürbare Belastung für das Team.

HERAUSFORDERUNG

Prokon betreibt kritische Infrastruktur. Ein einzelner Diagnosefehler kann Schäden im sechsstelligen Bereich verursachen. Wird eine Komponente falsch behandelt und ein Kranrückbau nötig, drohen Millionenkosten. An der Kompetenz liegt das nicht, denn die Techniker sind hochqualifiziert, viele haben Elektrotechnik oder Maschinenbau studiert. Die eigentliche Hürde ist die Komplexität. 20 Anlagengenerationen bedeuten 20 verschiedene Systeme, jedes mit eigenen Fehlercodes, eigenen Herstellern und eigenen Handbüchern. Tritt ein Fehlercode auf, fahren Techniker hunderte Kilometer und klettern bei älteren Anlagen ohne Aufzug hoch. Vor Ort entscheiden sie dann oft mit einer unvollständigen ersten Diagnose, und das macht eine zweite oder dritte Anfahrt nötig. Standard-KI-Lösungen helfen hier nicht weiter. Sie geben unsichere Antworten, führen keine mehrstufigen Analysen durch und halluzinieren im Zweifel. In kritischer Infrastruktur ist das keine Option.

ZIELE

Prokon verfolgte drei Ziele. Das erste waren weniger Anfahrten durch bessere Vorbereitung, denn jede vermiedene Fahrt spart Kosten, erhöht die Verfügbarkeit und schont das Team. Das zweite waren verlässliche Diagnosen ohne Halluzinationen, bei denen das System offen sagt, was es nicht weiß, statt zu raten. Das dritte war die Skalierung über Ländergrenzen. Polen, Finnland und Spanien stehen auf der Roadmap, und die Lösung muss dort mehrsprachig funktionieren, ohne jedes Mal neu implementiert zu werden.

LÖSUNG

octonomy arbeitet nicht mit simplen Keyword-Suchen oder vordefinierten Entscheidungsbäumen, sondern führt echte Analysen durch. Das System nimmt einen Fehlercode, gleicht ihn mit aktuellen Monitoring-Daten ab und durchsucht die relevanten Handbücher. Schritt für Schritt leitet es her, welche Komponenten betroffen sein könnten. Der Techniker bekommt keinen allgemeinen Hinweis wie „prüfen Sie die elektrische Anlage“, sondern einen strukturierten Diagnoseplan mit konkreten Messwerten und benötigtem Equipment. Die Integration in Microsoft Dynamics Field Service und die Monitoring-Systeme erfolgte über APIs, die bestehende IT-Infrastruktur blieb unverändert. Bei komplexen technischen Fragen erreicht octonomy 95 Prozent Genauigkeit, während Standard-LLMs oft unter 40 Prozent liegen. Die Lösung ist Made in Germany, DSGVO-konform und EU-AI-Act-ready.

UMSETZUNG

Prokon integrierte octonomy direkt in Microsoft Dynamics Field Service, als zentrale Plattform im Arbeitsablauf von Disponenten und Technikern. Als Datenquellen dienen die Monitoring-Systeme der Windkraftanlagen, die 2.600 Herstellerdokumente, die interne Prozessdokumentation und die Erfahrungswerte des Second-Level-Supports. Der Rollout startete mit einer Testgruppe im deutschen Markt, mit paralleler produktiver Testphase und iterativer Erweiterung. Zwei Wochen nach dem ersten Gespräch waren alle relevanten Stakeholder eingebunden. Die Machbarkeitsstudie mit messbaren Ergebnissen lag nach zehn Tagen vor, die erste produktive Nutzung folgte innerhalb von sechs Wochen.

ERGEBNIS

Die Anfahrten haben sich halbiert, die Problemlösung vor Ort ist 25 Prozent schneller, und die Anlagenverfügbarkeit ist messbar höher. Das technische Team arbeitet fokussierter, die Belastung sinkt spürbar. Das Feedback aus dem technischen Service übertraf die Erwartungen.

> „Mit octonomy reduzieren wir unsere Anfahrten um 50 Prozent. Das System gibt uns die Sicherheit, die wir in der kritischen Infrastruktur brauchen.“
>
> Dr. Yousef Farschtschi, Bereichsleiter IT und Digitales, Prokon

Das Problem war nicht fehlendes Wissen. Das Problem war, es im richtigen Moment am richtigen Ort zu haben.

Our Side — Wie wir es gebaut haben

Ausgangslage: Hilfe im Feld ohne digitale Diagnose

Prokon eG ist Deutschlands größte Energiegenossenschaft und betreibt unter anderem Windkraftanlagen. Deren technische Dokumentation ist sehr umfangreich. Für einen einzelnen Anlagentyp von Vestas liegen 2.641 Dokumente mit rund 50.443 Seiten vor, darunter Fehlercode-Verzeichnisse, Bauteilinformationen und elektrische Schaltpläne, die mit zahlreichen Querverweisen aufeinander verweisen. Insgesamt deckt die Lösung die Dokumentation von 20 verschiedenen Anlagengenerationen ab.

Unser Ansatz: ein Expertensystem direkt in der Arbeitsumgebung des Technikers

Bei diesem Umfang reicht ein einzelnes, monolithisches Chatbot-System nicht aus. Wir haben deshalb ein Multi-Agenten-System zum Troubleshooting von Windkraftanlagen entwickelt. Grundlage war zunächst die Vestas-Dokumentation, inzwischen sind weitere Anlagentypen hinzugekommen. Der Servicetechniker findet das System direkt in der Microsoft Field Service App auf dem Tablet, das er ohnehin mit auf die Anlage nimmt. Dort stellt er seine technischen Fragen, ohne ein zusätzliches Programm zu öffnen. Prokon läuft produktiv auf einer eigenen octonomy-Instanz.

Das System beantwortet ausschließlich Fragen zur Windkraftanlage. Nutzende können selbst festlegen, auf welche Informationen und Anlagentypen es zugreift. Ist nicht klar, für welchen Anlagentyp eine Anfrage gilt, fragt das System aktiv nach.

Jede Antwort folgt einem festen Aufbau. Zuerst kommt die Sofortmaßnahme, dann die betroffenen Bauteile, ein Diagnoseplan mit Zeitschätzung, die zu prüfenden Parameter, das empfohlene Werkzeug und die Sicherheitshinweise. So weiß der Techniker zuerst, was jetzt zu tun ist, und bekommt die vollständige technische Einordnung gleich dazu.

Prokon war der erste praktische Anwendungsfall für eine Fähigkeit der Plattform, die wir rekursive Werkzeugnutzung nennen. Das System beantwortet eine Anfrage nicht in einem einzelnen Schritt. Es verkettet mehrere Werkzeugaufrufe nacheinander, etwa von der Fehlerbeschreibung über die Bauteilliste bis zum Schaltplan, und stellt erst danach seine vollständige Antwort zusammen.

SYSTEM	ZIELGRUPPE	AUFGABE
Expertensystem zum Troubleshooting	Servicetechniker vor Ort	Beantwortet komplexe technische Fragen zu Windkraftanlagen, eingebettet in die Microsoft Field Service App

Der Aufbau erfolgte in drei Stufen, zuerst die Wissenskonsolidierung, dann die Teilautomatisierung und zuletzt die Vollautomatisierung. Damit sank die Zahl der nötigen Vor-Ort-Einsätze pro Störungsfall von durchschnittlich rund sechs auf rund drei.

Wie eine Diagnose aussieht: das Beispiel Fehler 181

Fehler 181 ist ein Turbinenfehler bei fehlerhaftem Rückmeldesignal während des Gierens im Uhrzeigersinn. Von der Meldung bis zur gefundenen Ursache führt ein fester Weg durch fünf Stufen.

1 FEHLER TRITT AUF

Fehler 181 „Feedback = _, yawing CW _“ · Turbinensteuerung meldet fehlerhaftes Rückmeldesignal

2 FEHLER IM LOGBUCH FINDEN

Typ / Kategorie	Error · Turbine
Signal Monitoring	S100A / S100B · 1 Sek. Scan
Kriterium	Feedback is not correct

Reaktion	Emergency Stop
	Die Turbine stoppt sofort alle Bewegungen, um Schäden durch unkontrollierte Gierbewegung zu verhindern
Wiederanlauf	Slow
	Kontrollierte statt abrupte Verzögerung, damit die mechanischen Systeme sicher herunterfahren können
Quittierung	Remote (SCADA)
	Der Fehler muss remote bestätigt werden, damit die Anlage nicht ohne menschliches Eingreifen neu startet

3 FEHLER ENTSCHLÜSSELN

- Sensor gestört oder getrennt
- Encoder- oder Resolver-Defekt
- Verkabelungsproblem
- Mechanischer Schlupf im Gierantrieb

Ergebnis: am wahrscheinlichsten ein elektrisches Problem.

4 VOM SIGNAL ZUM BAUTEIL

Signale sind nicht referenziert
Im Schaltplan steht kein Signalname — er muss erst aufgelöst werden

Abkürzungsverzeichnis
Projektspezifische Kürzel verbinden Signal und Bauteil

Bauteilliste
Komp.-Nr., Funktion, Ort, Seite und Spalte im Plan

5 IM SCHALTPLAN PRÜFEN

Kontaktor prüfen → K100B über Koordinate finden → nächstes Bauteil → **6× wiederholen**

Jede Planseite trägt eigene Metadaten, ein Koordinatensystem lokalisiert das Bauteil auf der Seite.

BRANCHE 05

Logistik

Die Logistik lebt von Regeln, die genau eingehalten werden müssen. Das nötige Wissen verteilt sich über viele Quellen, von Gefahrgutvorschriften über Prozess- und Speditionshandbücher bis zu Zoll- und Compliance-Vorgaben. Diese Regelwerke umfassen oft tausende Seiten und ändern sich laufend. Eine einzelne Auskunft kann mehrere Quellen gleichzeitig betreffen, und wer eine davon übersieht, riskiert eine unvollständige Antwort. Bei sicherheitsrelevanten Vorschriften wiegt ein Fehler schwer. Hinzu kommt die internationale Belegschaft, die verteilt über viele Standorte arbeitet und Informationen in unterschiedlichen Sprachen braucht. Jede manuelle Suche im richtigen Dokument kostet dabei Minuten, die im Tagesgeschäft fehlen. octonomy versteht verteilte Regelwerke, führt Wissen aus mehreren Quellen zusammen und liefert Antworten mit Kapitel- und Quellenangabe in Sekunden.

07 — Emons Spedition zeigt, wie sich dieses Regelwerk in Sekunden zugänglich machen lässt.

Emons Spedition

95 %+	3 Wochen	~100	20+
Antwortqualität	bis Go-live	unterstützte Sprachen	Handbücher integriert



„octonomy hat unser Wissen, das vorher statisch in Datenbanken lag, auf eine natürliche Kommunikationsweise greifbar gemacht."

Tilo Hergarten, Managing Director @ Emons | DIGITAL GmbH

AUF EINEN BLICK

Das Unternehmen	International tätiges Logistikunternehmen aus Köln, seit 1928, mit Transport- und Logistikdienstleistungen über Straße, Schiene, Luft und See. Eingesetzt im internen Wissensmanagement.
Ausgangslage	Ein wachsendes Netzwerk an Depots, eine international aufgestellte Belegschaft und steigende regulatorische Komplexität. Das relevante Wissen ist über mehr als 20 Handbücher mit tausenden Seiten Regelwerk verteilt.
Was Emons Spedition gemacht hat	Erweiterung des internen Wissensmanagements um die octonomy KI-Expertenlösung. Der Agent versteht Fragen in natürlicher Sprache, führt Wissen aus mehreren Handbüchern zusammen und liefert Antworten mit Kapitel- und Quellenangabe in Sekunden, direkt im Intranet.
Ergebnis	Über 95 Prozent Antwortqualität, verifiziert durch Ground-Truth-Testing. Go-Live in drei Wochen. Weniger Fehler bei Gefahrgutregelungen, schnellere Einarbeitung neuer Mitarbeiter und ein einheitlicher Wissensstand im Netzwerk.

VIDEOS UND MATERIAL

CASE STUDY

SUCCESS STORY

Die Case Study

AUSGANGSLAGE

Die Emons Spedition ist ein international tätiges Logistikunternehmen mit Sitz in Köln. Seit 1928 bietet das familiengeführte, mittelständische Unternehmen umfassende Transport- und Logistikdienstleistungen über Straße, Schiene, Luft und See. Tausende Mitarbeiter in zahlreichen Depots deutschlandweit arbeiten täglich mit einem der komplexesten Regelwerke der Branche, über 20 operative Handbücher zu Gefahrgut, Prozessen, Spedition und mehr.

Das Netzwerk an Depots wächst, die Belegschaft ist international aufgestellt, und die regulatorische Komplexität steigt. Damit wächst der Druck auf jeden einzelnen Mitarbeiter. Die richtigen Informationen müssen schnell, vollständig und fehlerfrei verfügbar sein.

HERAUSFORDERUNG

Das operative Regelwerk umfasst unter anderem das Speditionshandbuch, das Fahrerhandbuch und das Gefahrguthandbuch, dazu Prozessmanagement- und Packmittelhandbücher sowie ESG- und Compliance-Dokumente. Will ein Mitarbeiter wissen, welche Höchstmenge für eine bestimmte UN-Nummer gilt, muss er zunächst das richtige Handbuch identifizieren, die korrekte Tabelle finden, die aktuelle Version sicherstellen und die Information korrekt interpretieren. Ein solcher Vorgang dauerte bis zu 15 Minuten pro Anfrage und funktionierte besonders schlecht, wenn die Antwort über mehrere Handbücher verteilt war. Wer etwa wissen will, wie mit beschädigtem Gefahrgut umzugehen ist, muss sowohl das Gefahrgut- als auch das Fahrerhandbuch kennen. Wer eines übersieht, riskiert eine unvollständige und im Zweifel sicherheitskritische Antwort. Hinzu kommt die internationale Belegschaft. Informationen, die nur auf Deutsch verfügbar sind, erreichen nicht alle Mitarbeiter gleich gut.

ZIELE

Die Emons Spedition hat den Fokus auf drei Ziele gelegt. Das erste ist die Entlastung im Tagesgeschäft. Mitarbeitende sollen keine Arbeitszeit mehr mit dem Suchen in langen PDFs verbringen, sondern Routinefragen zu Gefahrgut, Prozessen und Sendungsvorschriften sofort und zuverlässig beantwortet bekommen. Das zweite ist die regulatorische Präzision. Gerade bei Gefahrgutvorschriften mit UN-Nummern, ADR-Klassen und Mengengrenzen sind Fehler keine Option, und die Antworten müssen nachvollziehbar, quellenbasiert und fehlerfrei sein. Das dritte ist ein einheitlicher Wissensstand im Netzwerk. Ob ein Mitarbeiter in Köln, in einem Regionaldepot oder international arbeitet, alle sollen denselben Zugang zu denselben Informationen haben, in ihrer jeweiligen Sprache.

LÖSUNG

octonomy versteht die über 20 operativen Handbücher aus SharePoint, erkennt Zusammenhänge zwischen Dokumenten und liefert Antworten mit präziser Kapitel- und Versionsnummer. Fragt ein Mitarbeiter, wie mit beschädigtem Gefahrgut umzugehen ist, zieht der Agent die relevanten Passagen aus Gefahrgut- und Fahrerhandbuch zusammen und liefert eine vollständige, handlungsrelevante Antwort in einem Schritt. Eine manuelle Tabellensuche zu einer UN-Nummer dauerte bis zu 15 Minuten, mit octonomy dauert sie Sekunden, inklusive Quellenangabe mit Versionsnummer. Das System unterstützt rund 100 Sprachen, primär Deutsch und Englisch, und ist als Chat-Widget direkt im Emons-Intranet eingebettet, in der Sie-Form und im Corporate Design von Emons, ohne neues Tool und ohne neuen Login.

UMSETZUNG

Die Emons Spedition und octonomy sind pragmatisch gestartet, mit dem Ziel, Anforderungen klar zu definieren, das System aufzubauen, zu testen und live zu stellen. Vom ersten gemeinsamen Gespräch bis zum produktiven Einsatz vergingen drei Wochen. In den ersten zwei Tagen wurden Projektplan und Verantwortlichkeiten festgelegt. In Woche eins folgte der Aufbau des Expertensystems, mit dem Erstellen der Wissensbasis, der Integration der Handbücher und der Definition der KI-Richtlinien. Woche zwei stand im Zeichen der Testphase, mit Fokus auf Funktionalität, Usability und Ground-Truth-QA. In Woche drei wurden Key User geschult und das System in der Produktivumgebung live gestellt.

ERGEBNIS

Die Antwortqualität liegt bei über 95 Prozent, verifiziert durch Ground-Truth-Testing. Der Go-Live war nach drei Wochen erreicht. Es gibt weniger Fehler bei Gefahrgutregelungen, eine schnellere Einarbeitung neuer Mitarbeiter und einen einheitlichen Wissensstand im gesamten Netzwerk.

VOM ERSTEN GESPRÄCH ZUM GO-LIVE — DREI WOCHEN

TAG 1–2	WOCHE 1	WOCHE 2	WOCHE 3
Setup	Aufbau	Testphase	Go-Live
Projektplan und Verantwortlichkeiten	Wissensbasis, Integration der Handbücher, KI-Richtlinien	Funktionalität, Usability, Ground-Truth-QA	Key-User-Schulung, Produktivumgebung

Our Side — Wie wir es gebaut haben

Ausgangslage: Wissen, das in zwei Systemen gleichzeitig steckt

Die Emons Spedition ist ein Speditions- und Logistikunternehmen und pflegt einen erheblichen Teil seines internen Wissens redaktionell in einem internen Intranet-System. Dazu kommen rund 400 Prozessdiagramme in einem Prozessmodellierungs-Tool, von denen etwa 100 bis 120 direkt im Intranet verlinkt sind. Für einen Chatbot war dieses Wissen bislang nicht direkt zugänglich. Die Prozessdiagramme sind im Intranet nur als Embed-Link eingebunden, und ein Klick öffnet ein separates Fenster mit einer reinen Pixelgrafik. Eine gewöhnliche Anbindung des Intranet-Systems sieht dabei nur den Link, nicht das Bild selbst.

Unser Ansatz: EmonsGPT in zwei Schritten

Wir haben für die Emons Spedition EmonsGPT gebaut, einen internen Wissens-Chatbot. Er beantwortet Mitarbeitenden Fragen zu Prozessen und Regelwerken direkt im Chat. Den ersten Schritt, die Anbindung der internen Prozesse und Regelwerke auf Basis von SharePoint, haben wir seit dem 23. März 2026 live. Seit dem Go-Live zählen wir mit Stand August 2026 über 800 Konversationen. 76,2 Prozent der Fragen konnte EmonsGPT direkt beantworten. Weitere 22,6 Prozent betrafen Themen, die noch nicht Teil der Wissensbasis waren und die der Agent korrekt als solche erkannt hat. In Summe hat er damit in 98,8 Prozent der Fälle richtig gehandelt, entweder mit der richtigen Antwort oder mit der richtigen Einschätzung, dass er die Antwort noch nicht kennt.

>800*	76,2 %*	98,8 %*
Konversationen seit Go-Live	direkt beantwortet	richtig gehandelt

* Alle Kennzahlen mit Stand August 2026.

DER ZWEI-STUFEN-AUFBAU VON EMONSGPT

BAUSTEIN 1

Interne Prozesse und Regelwerke

Über 20 operative Handbücher aus SharePoint, Antworten mit Kapitel- und Versionsnummer

- Live seit 23.03.2026, aktiv genutzt

BAUSTEIN 2

Das vollständige Intranet

Intranet-System über Site-to-Site-VPN, dazu rund 400 Prozessdiagramme, Format-Entscheidung in Arbeit

- Technisch vorbereitet, Rollout mit Zielrichtung Jahresende

Der zweite Schritt ist die vollständige Erschließung des Intranets, über einzelne Dokumente hinaus. Die technische Anbindung der Intranet-Seiten steht bereits. Dafür war eine Site-to-Site-VPN-Verbindung zwischen unserem Rechenzentrum und der Emons Spedition notwendig, weil das Intranet nicht öffentlich aus dem Internet erreichbar ist. Aktuell bereinigt die Emons Spedition die Inhalte des Intranets selbst. Sobald das abgeschlossen ist, lassen sich die Intranet-Daten als weiterer Teil der Wissensbasis nutzen.

Wie wir das Prozessdiagramm-Problem angehen

Für die Prozessdiagramme haben wir mehrere Formate verglichen. Das aktuelle Pixelbild lässt sich per Bilderkennung interpretieren, allerdings ungenau. Über die API des Prozessmodellierungs-Tools ist zusätzlich ein JSON-Format verfügbar, das Abhängigkeiten und Entscheidungspunkte explizit enthält. Aus unserer Sicht ist das voraussichtlich das aussagekräftigste Format. BPMN 2.0 XML ist inhaltlich vergleichbar tief und für Menschen besser lesbar. SVG ist rein deskriptiv und für die KI-Verarbeitung am wenigsten hilfreich. Unsere Empfehlung ist eine einmalige, statische Einlese in die Wissensbasis, ergänzt um einen optionalen Test mit unserem bestehenden Website-Search-Feature.

Ein Vorteil auf Emons-Seite ist die klare Benennung. Jeder modellierte Prozess trägt einen eindeutigen Namen, ohne Duplikate und identisch zum im Intranet verlinkten Titel. Das macht die Zuordnung durch die KI grundsätzlich unproblematisch, sobald das Format-Thema gelöst ist.

Datenschutz und Zugriffsrechte

Der an uns vergebene Zugriffstoken soll ausschließlich mit den für EmonsGPT vorgesehenen Rechten ausgestattet sein, ohne ein 1:1-Mapping individueller Nutzerrechte, wie es bei anderen Copilot-Ansätzen üblich ist. Inhalte mit eingeschränkter Sichtbarkeit, etwa nur für die Geschäftsführung, dürfen nicht ungefiltert in die allgemein zugängliche Wissensbasis gelangen. Unsere Empfehlung ist deshalb, Inhalte nach Sensibilität in getrennten Ordnerstrukturen zu organisieren, statt sie nachträglich zu filtern.

Stand heute

EmonsGPT ist mit dem ersten Baustein, den internen Prozessen und Regelwerken, seit Ende März 2026 live und wird aktiv genutzt. Der zweite, größere Baustein, das vollständige Intranet inklusive der Prozessdiagramme, ist technisch vorbereitet und wartet auf die von der Emons Spedition selbst durchgeführte Datenbereinigung. Emons bevorzugt bewusst einen vollständigen Rollout dieses zweiten Bausteins auf einmal, aus Gründen der Kohärenz und Kommunikation, im Rahmen eines bereits vereinbarten Daten-Gesamtpakets mit Zielrichtung Ende des laufenden Jahres. Mittelfristig plant die Emons Spedition zudem eine Migration des Intranets auf eine SharePoint-Basis. Nach unserer Einschätzung lässt sich das grundsätzlich unproblematisch anbinden.

Customer Side — Interview mit Emons Spedition

Auszug aus dem Gespräch mit Tilo Hergarten, Managing Director, Emons | DIGITAL GmbH.

Was war der Auslöser, sich mit octonomy zu beschäftigen?

> „Wir sind fast 100 Jahre alt, da baut sich viel auf, was man im Gepäck hat und trotzdem im Alltag nutzbar machen muss. Für uns war der Punkt, dass wir einen Hebel brauchten, mit dem wir dieses Wissen auf ein neues System transferieren können, damit es für jeden Nutzer möglichst schnell zugänglich ist.“
>
> Tilo Hergarten, Managing Director @ Emons | DIGITAL GmbH

Wie haben die Mitarbeitenden reagiert?

> „Unsere Mitarbeiter haben auf die Veränderung grundsätzlich sehr positiv reagiert. Als es funktioniert hat, kamen einige und sagten: Aber das muss noch rein, und das muss noch rein. Die Neugier ist da, die Nutzung ist da, die Qualität ist da. Insofern wünschen sich die Mitarbeiter, dass es weitergeht, und das ist ein sehr positives Zeichen für die Akzeptanz.“
>
> Tilo Hergarten, Managing Director @ Emons | DIGITAL GmbH

Gab es einen Moment, in dem klar wurde, dass octonomy wirklich funktioniert?

> „Man gibt eine Frage ein und man bekommt das Ergebnis und man weiß, dass es stimmt. Und das ist natürlich dann der Moment, wo man weiß: Es hat ganz gut funktioniert.“
>
> Tilo Hergarten, Managing Director @ Emons | DIGITAL GmbH

Wissen, das im Unternehmen steckt, nützt nichts, solange es nicht abrufbar ist.



Software

In Software- und Technologieunternehmen liegt das Wissen über einen Kunden in vielen Systemen gleichzeitig. Es steckt in Support-Tickets, CRM-Notizen, Dokumentationsportalen, internen Wikis, Code-Repositories und der Zeiterfassung. Was fehlt, ist der eine Blick auf den Account. Wer eine Upsell-Chance oder ein Kündigungsrisiko erkennen will, muss die einzelnen Teile von Hand zusammensuchen, meist kurz vor dem Kundengespräch. octonomy fragt die Systeme parallel ab, trennt Chancen konsequent von Risiken und macht jeden Befund auf Quelle und Datum rückverfolgbar.

08 — ELAINE technologies zeigt, wie ein Agent über neun Systeme hinweg schlussfolgert.

ELAINE technologies

Elaine
an Entirely company

9	25	20 Sek.	4
Systeme parallel abgefragt	Analyse-Muster	statt 1,5 Std. Recherche	Intelligence-Domänen

AUF EINEN BLICK

Das Unternehmen	Deutsches Customer-Engagement-Softwareunternehmen, Marke „Entirely“, mit Konzernkunden im Bestandskundenportfolio. Eingesetzt im Customer Success zur Account-Analyse.
Ausgangslage	Das Customer-Success-Team betreut über 50 Bestandskunden, darunter auch Konzernkunden. Entscheidende Kunden-Informationen lagen verstreut über sechs getrennte Systeme, unter anderem in rund 70 Support-Tickets pro Woche.
Was ELAINE gemacht hat	Ein konversationeller KI-Agent namens „Elaine“ bestätigt den Account, zieht Daten parallel aus bis zu neun Systemen, trennt Chancen von Risiken und schreibt die Zusammenfassung auf Wunsch ins CRM.
Ergebnis	Antwortzeit je nach Komplexität 20 Sekunden bis 8 Minuten statt 1,5 bis 2,5 Stunden manueller Recherche. Live seit April 2026, ein zweiter Anwendungsfall ist beauftragt.

Our Side — Wie wir es gebaut haben

Ausgangslage: sechs Systeme, kein Gesamtbild

ELAINE technologies (Marke „Entirely“) ist ein deutsches Customer-Engagement-Softwareunternehmen. Das eigene Customer-Success-Team betreut über 50 Bestandskunden, darunter auch Konzernkunden, und bearbeitet dabei rund 70 Support-Tickets pro Woche.

Die entscheidende Information lag verstreut über sechs getrennte Systeme, darunter Jira, Confluence und HubSpot. Jedes System beantwortete eine Teilfrage. Offene Tickets standen an einer Stelle, Customizing-Details an einer zweiten, die eingesetzte Produktversion an einer dritten. Was fehlte, war der eine Blick auf den Account. Das Team stellte ihn vor jedem Kundengespräch von Hand her, mit 1,5 bis 2,5 Stunden Recherche pro Analyse.

Das Zielbild: eine 360-Grad-Sicht auf den Kunden

Intern lief das Vorhaben unter dem Namen Customer Insights Agent, auch 360-Grad-Agent genannt. Der Name beschreibt die Absicht genau. Gewollt war ein internes Assistenzsystem für Customer Success und Kundenservice, das die verstreuten Informationen an einer Stelle zusammenführt. Interne wie externe Anfragen sollten sich damit schneller beantworten lassen. Das vollständige Bild eines Kunden sollte zugleich Upselling-Potenziale früher sichtbar machen und Bedarfe über die Lösungen der Unternehmensgruppe hinweg verständlich machen.

Der Anspruch reichte über den ersten Anwendungsfall hinaus. Ein weiteres isoliertes Tool war ausdrücklich nicht gewollt. Gefragt war eine Plattform, auf der sich KI schrittweise im Unternehmen verankern lässt, später skalierbar auf weitere Bereiche der Entirely-Gruppe und auf eine mögliche Automatisierung von 24/7-Supportprozessen.

Entsprechend explizit war die Anforderung, mit der ELAINE zu uns kam. Kein simpler Chatbot. Gesucht war eine Lösung, die über die gesamte Tool-Landschaft hinweg schlussfolgert und dem Customer Success Manager vor dem Gespräch eine strukturierte, quellenzitierte Antwort in die Hand gibt.



Unser Ansatz: ein Orchestrator, viele parallele Sub-Agenten

Unsere Lösung war ein konversationeller KI-Agent, passend zum Namen des Kunden „Elaine“ getauft. Der Agent spricht direkt mit dem Customer Success Manager. Fragt dieser beispielsweise nach einem beliebigen Kunden-Account, läuft im Hintergrund ein mehrstufiger Prozess: Der Agent bestätigt zunächst den Account, per Name oder Account-Nummer, und löst dabei automatisch Namensvarianten und Sub-Marken-Benennungen auf. Dann zieht er Daten parallel aus bis zu neun Systemen: CRM, Ticketing, Zeiterfassung, Finanzdaten, Code-Repository, Produkt- und Nutzungssystem, Dokumentationsportal, Wiki und Release-Historie. Sechs davon waren vorher Handarbeit. Anschließend trennt er Chancen konsequent von Risiken, nie vermischt, mit jedem Befund rückverfolgbar auf Quelle und Datum. Auf Wunsch schreibt er die Zusammenfassung direkt als Notiz ins Kundenmanagement-System, damit nach dem Gespräch nichts verloren geht.

Unter der Haube steckt dafür eine Multi-Agenten-Architektur statt eines einzelnen, monolithischen Systems:

AGENT	ROLLE	AUFGABE
Orchestrator „Elaine“	Gesprächspartner des Customer Success Managers	Nimmt die Frage auf, bestätigt den Account, matcht sie gegen ein vorgebautes Analyse-Muster, delegiert an Sub-Agenten, führt die Ergebnisse zu einer strukturierten Antwort zusammen
Token Collector	Vorbereitender Sub-Agent	Ermittelt über leichte, reine ID-Abfragen die vollständige Liste der relevanten Datenseiten, bevor die eigentliche Analyse beginnt
Analyse-Worker (parallel)	Ausführende Sub-Agenten	Je ein Worker pro Datenseite, für Aggregation, Themenerkennung oder detaillierte Tabellenauswertung
Account-Health-Agent	Zweiter, separat beauftragter Anwendungsfall	Verwandter Anwendungsfall zur Kundenzufriedenheit, im Anschluss an Elaine beauftragt

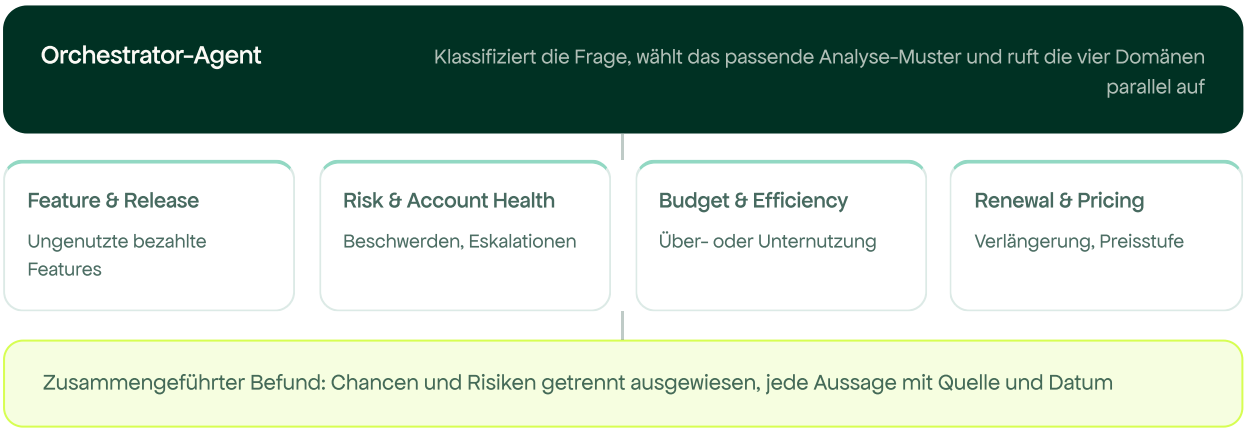
Der entscheidende Designpunkt dabei: Jeder Analyse-Worker sieht immer nur einen begrenzten Ausschnitt der Daten, eine Seite. Das hält das System schnell und verhindert, dass es mit wachsendem Datenvolumen langsamer wird, die Arbeit wird über Datenseiten parallelisiert statt sequenziell über den ganzen Account abgearbeitet. So bekommen selbst Accounts mit hunderten offenen Tickets eine Antwort in Sekunden statt Minuten.

Vier Intelligence-Domänen

Die Fragen, die der Agent beantwortet, lassen sich in vier Domänen einordnen, abgedeckt über 25 unterschiedliche Frage-Muster.

DOMÄNE	TYPISCHE FRAGE
Feature & Release Intelligence	Hängt dieser Account bei Features hinterher, für die er bezahlt, oder würde er von einem jüngsten Release profitieren?
Risk & Account Health	Zeigt dieser Account Warnsignale wie Beschwerden, Eskalationen oder SLA-Verletzungen?
Budget & Service Efficiency	Über- oder unternutzt der Account seine vertraglich vereinbarten Service-Stunden?
Renewal & Pricing Intelligence	Steht eine Verlängerung an, und rechtfertigt die aktuelle Nutzung eine andere Preisstufe?

SO STEUERT DER ORCHESTRATOR



Wichtig dabei: Der Agent erfindet seine Suchlogik nicht selbst. Bevor überhaupt ein System abgefragt wird, gleicht der Orchestrator die Frage gegen einen kuratierten Katalog vorgebauter Analyse-Muster ab und wählt daraus ein vorab geprüftes, von Menschen verfasstes Set an Abfrage-Vorlagen, statt sich eine eigene Suchstrategie zurechtzulegen.

Regelwerk gegen Halluzinationen

Mehrere Absicherungen laufen dabei konsequent mit. Ist ein Feld in der Quellsystem-Antwort leer oder fehlt es, sagt der Agent das explizit, statt einen plausibel klingenden Wert einzusetzen.

Für jeden Datentyp gilt genau ein System als maßgebliche Quelle. Etwa ist das Nutzungssystem die einzige Quelle dafür, welche Features ein Account nutzt. Der Agent darf diese Informationen nicht aus allgemeinem Modellwissen beantworten.

Ist der jüngste Datenpunkt zu einem Thema älter als sechs Monate, weist der Agent proaktiv darauf hin, statt veraltete Daten als aktuell zu präsentieren. Und jede berechnete Zahl, Stunden, Umsatz, Prozente, kommt mit der Anzahl der zugrunde liegenden Datensätze zurück, sodass der Customer Success Manager sie prüfen kann, bevor er sie an einen Kunden weitergibt.

Sicherheit und Compliance

Die Plattform arbeitet mit ISO 27001/27701-zertifiziertem Betrieb (SOC 2 im Geltungsbereich), ausschließlich EU-Datenverarbeitung, einschließlich des zugrunde liegenden Sprachmodells, grenzüberschreitende Datenflüsse sind über Standardvertragsklauseln und ein abgeschlossenes Transfer Impact Assessment abgedeckt, und der Support erfolgt ausschließlich durch EU-basiertes Personal.

Stand heute

Der Agent ist seit April 2026 live im produktiven Einsatz und wird in der Anfangszeit noch eng begleitet und laufend weiter verfeinert. Die Antwortzeit liegt je nach Komplexität zwischen 20 Sekunden und 8 Minuten, verglichen mit den vorher üblichen 1,5 bis 2,5 Stunden manueller Recherche. ELAINE technologies hat inzwischen den Account-Health-Agenten als zweiten, verwandten Anwendungsfall beauftragt, ein Zeichen dafür, dass ein erfolgreicher erster Anwendungsfall auch innerhalb eines bestehenden Kunden weiteres Vertrauen und weitere Zusammenarbeit eröffnet.

PRINZIP

Jeder Analyse-Worker sieht nur eine Datenseite. Deshalb wird das System mit wachsendem Datenvolumen nicht langsamer.

BRANCHE 07

Klima- und Heiztechnik

Die Heiz- und Klimatechnik lässt wenig Spielraum für ungenaue Antworten. Jede Anfrage trifft auf hunderte Gerätevarianten, jede mit eigenen Schaltplänen und Fehlercodes. Ohne Systemabgleich beginnt jede Auskunft mit einer manuellen Suche. In der Heizperiode kommt ein zweites Problem dazu. Das Anrufvolumen steigt deutlich, die Teamgröße bleibt gleich. Bei sicherheitsrelevanten Fällen stößt ein Standard-Bot schnell an seine Grenze und kann am Ende nur einen Techniker empfehlen. Hinzu kommt das Haftungsrisiko, denn ohne Quellenangabe lässt sich eine Antwort im Schadensfall nicht belegen. octonomy liest Installationshandbücher, Schaltpläne und Fehlercode-Tabellen bildhaft. Die Diagnosen kommen mit konkretem Klemmen-, Spannungs- und Bauteilbezug und verweisen auf die exakte Handbuchseite.

09 — Stiebel Eltron Schweiz zeigt, wie technischer Service davon entlastet wird.

Stiebel Eltron Schweiz

STIEBEL ELTRON

~45	~700	3	407
Techniker im Feld	technische PDFs	Sprachen DE/FR/IT	ausgewertete Konversationen



„Dieses Projekt war sehr gute Arbeit. Ich bin überzeugt, wenn wir jetzt unsere Wissensdatenbank weiter pflegen, wird das hier ein Produkt sein, das in Zukunft keiner mehr missen möchte!“

Reinhard Unger, IT-Leiter, Stiebel Eltron Schweiz

AUF EINEN BLICK

Das Unternehmen	Schweizer Landesgesellschaft des Herstellers von Wärmepumpen und Heizsystemen, mit eigenem technischen Service. Eingesetzt im Field Service für Wartungs- und Servicetechniker.
Ausgangslage	Recherchen liefen über eine Excel-Liste. Die Basis: Ein Index auf rund 300 PDF-Dokumente in einer lokalen Dateiablage. Dauer: Rund 10 Minuten pro Anfrage. Ein Großteil des Erfahrungswissens steckte in den Köpfen von nur drei Anlagenspezialisten.
Was Stiebel Eltron gemacht hat	Durch SELIA, einem KI-Assistenten auf der octonomy-Plattform, wird die technische Wissensbasis semantisch durchsuchbar — hinter dem eigenen Single Sign-On, mit Quellenangabe und Seitenzahl.
Ergebnis	Live seit August 2026, mit stufenweisem Feld-Rollout. Ziel ist eine Reduktion der Recherchezeit von rund 10 Minuten auf rund 1 Minute pro Fall.

Our Side — Wie wir es gebaut haben

Ausgangslage: Wissen, das mit der Pensionierung droht zu verschwinden

Stiebel Eltron Schweiz betreibt einen eigenen technischen Service für Wärmepumpen und Heizsysteme, mit rund 45 Service- und Wartungstechnikern im Feld. Mit anstehenden Pensionierungen drohte spürbarer Wissensverlust. Wissenskonservierung war deshalb von Anfang an der eigentliche Treiber des Projekts, nicht nur Zeitersparnis.

Unsere Lösung: SELIA

Wir haben SELIA gebaut, einen KI-Assistenten auf der octonomy-Plattform. Er macht die gesamte technische Wissensbasis der Schweizer Landesgesellschaft semantisch durchsuchbar.

Ein Techniker stellt seine Frage in natürlicher Sprache. Das läuft über ein Web-Frontend hinter dem Stiebel-eigenen Single Sign-On. Es geht um ein Fehlerbild, ein Ersatzteil oder das Zusammenspiel von Regler und Anlage.

Er erhält eine quellenbasierte Antwort mit genauer Dokumentangabe und Seitenzahl. So lässt sich jede Auskunft im Original nachschlagen.

Ist die Gerätebezeichnung nicht eindeutig, fragt SELIA aktiv nach. Das passiert etwa dann, wenn mehrere Anlagengenerationen ähnlich heißen. Die Antworten kommen in der Sprache des Nutzers: Deutsch, Französisch oder Italienisch.

Die Wissensbasis liegt in einem Azure-Blob-Container beim Kunden. Das ist ein Cloud-Speicher von Microsoft Azure, in dem die PDF-Dokumentation liegt. Wir greifen darauf über einen Service Principal zu.

Im produktiven Einsatz sind das rund 700 PDFs, etwa 2 GiB. Eine Excel-Masterliste vom Kunden steuert, welche Dokumente dazugehören.

Darüber liegt ein Metadaten-Layer. Jedes Dokument ist eindeutig gelabelt: nach Ordnerstruktur, Dokumentcode und Gerätebezeichnung, zum Beispiel „WPL-A 07.2 HK 230 Premium“. Zusätzlich gibt es Lookup-Datenbanken, etwa einen Controller-Lookup, der Maschinen den passenden Reglern zuordnet.

Dieses Referenzwissen haben wir nicht einfach als normale Dokumente eingespielt. Wir haben es als eigene, strukturierte Datenquelle angebunden. So kann SELIA solche Zuordnungen zuverlässig beantworten, statt sie aus Fließtext herauslesen zu müssen.

Die Qualität haben wir von Anfang an gesichert, über Ground-Truth-Frage-Antwort-Paare. Der Kunde hat sie geliefert. Sie erden den Agenten und dienen als feste Messlatte für jede Testrunde.

Ein kuratierter Testerkreis hat die Antworten geprüft, mit Muttersprachlern für Französisch und Italienisch. Insgesamt entstanden so 407 einzeln ausgewertete Konversationen, bevor SELIA live ging.

BAUSTEIN	TECHNISCHE UMSETZUNG
Web-Frontend	Hinter Stiebels SSO (Azure AD), mehrsprachig DE/FR/IT
Wissensbasis	Azure-BLOB-Container des Kunden, rund 700 PDFs, Zugriff über Service Principal
Metadaten-Layer	Kanonische Produktliste, Nomenklatur-Schlüssel, Dokumenttyp-Prioritäten, Geräte-Zuordnung
Gegenfragen-Flow	Aktive Rückfrage bei nicht eindeutiger Gerätebezeichnung
Infrastruktur	AWS Frankfurt (eu-central-1), LLM primär Claude über AWS Bedrock (EU), sekundär OpenAI über Azure OpenAI (EU)
Insights Board	Dashboards für Executive-, Operations- und Optimierungssicht, mit Resolution Rate und Cost Savings

Der Anwendungsfall im Alltag

Im Kern ersetzt SELIA für die Techniker den alten Weg: erst im Excel-Index suchen, dann das richtige PDF finden. Der Techniker muss kein Dokument mehr identifizieren und darin nach der passenden Stelle suchen. Er bekommt die Antwort direkt, mit Verweis auf Dokument und Seite als Beleg.

Das Insights Board macht den Nutzen zusätzlich sichtbar. Es zeigt, wie viele Anfragen SELIA eigenständig löst und wo Wissenslücken in der Wissensbasis liegen. Daraus leitet es konkrete Handlungsempfehlungen ab, etwa welches Dokument als Nächstes ergänzt oder überarbeitet werden sollte.

Für Fälle, die SELIA nicht abschließend beantworten kann, ist eine automatische Eskalation per Mail an das Spezialteam bereits konzipiert. Sie kommt in einer späteren Ausbaustufe.

Status Quo August 2026

SELIA ist seit August 2026 live, vollständig auf der Produktionsumgebung, hinter Stiebels SSO.

Der stufenweise Feld-Rollout an die Techniker ist in Planung. Zunächst mit Produktmanagement und einem weiteren Standort. Eine zweite Stufe folgt rund einen Monat später.

Das Ziel: Recherchezeit von rund 10 Minuten auf rund 1 Minute pro Fall senken. Belastbare, gemessene Betriebskennzahlen aus dem echten Feldeinsatz liefert das Insights Board nach den ersten Live-Wochen.

Phase 2 ist bereits unterzeichnet: ein Ops-Agent plus ein weiteres Expertenteam. Sie startet, sobald Phase 1 erfolgreich abgeschlossen ist.

Die Schweiz ist bewusst als Pilot gewählt, mit Blick auf eine mögliche Skalierung nach Deutschland mit rund zehnfachem Volumen.

Problem solved. Punkt? Nein, Komma.

Ein gelöstes Problem klingt nach einem Schlusspunkt. In diesem Heft ist es fast immer ein Komma.

hepster hat mit Produktinformationen angefangen und automatisiert heute Standardanfragen bis ins Policy-System. BORA ist intern gestartet und baut inzwischen den dritten Einsatzbereich auf. Die ANWR GROUP begann mit einem klar abgegrenzten POC und will ANNI zum einzigen Einstiegspunkt für ihre Händler machen. Kein Fall in diesem Heft endet dort, wo er angefangen hat.

Genau das ist das Muster: klein anfangen, messen, und erst dann den nächsten Schritt gehen. Keines der Unternehmen hat sein Team ersetzt, keines hat einen Schnellschuss gemacht. Die Marktdaten aus der Einleitung zeigen, wie selten das ist: 95 Prozent der GenAI-Piloten liefern keinen messbaren Mehrwert. Der Unterschied liegt in der Antwortqualität des eingesetzten Systems. 96 Prozent gegenüber 26 Prozent entscheiden darüber, ob ein System im Echtbetrieb funktioniert oder nicht.

Da das Muster funktioniert, ist auch dieses Heft nicht fertig. Neue Fälle kommen laufend dazu, im gleichen Aufbau: Ausgangslage, Herausforderung, Ziele, Lösung, Umsetzung, Ergebnis, dazu das Interview mit dem Kunden und die Einschätzung des octonomy-Teams.

Die Unternehmen in diesem Buch haben diesen Unterschied früh erkannt. Deichmann, Atlas Copco, Berner, HAVI und weitere folgen in der nächsten Ausgabe.

Einen Punkt gibt es hier nicht. Nur Kommas.

Wie octonomy arbeitet: Verstehen. Antworten. Handeln.

Alle Fälle in diesem Heft basieren auf derselben Plattform. Drei Säulen arbeiten dabei als Gesamtsystem zusammen, das unterscheidet octonomy von Chatbots und Ticket-Automatisierungen.

Verstehen

octonomy liest die vorhandene Dokumentation vollständig, auch das, was generische KI nicht verarbeitet: Schaltpläne, Kompatibilitätstabellen, annotierte Zeichnungen, AVB-Klauseln, strukturierte Produktübersichten. Jede Seite wird als Bild gelesen, Struktur und Kontext bleiben erhalten. Nichts geht verloren.

Antworten

Ein Supervisor-Agent koordiniert ein Team spezialisierter KI-Agenten und führt Informationen aus mehreren Quellen und Systemen zusammen. Das Ergebnis ist eine geprüfte, quellenbasierte Antwort mit 96 Prozent Genauigkeit, unabhängig validiert durch die Tolly Group. Wenn eine Information fehlt, wird das transparent gemacht, statt zu raten. Keine Halluzinationen.

Handeln

octonomy bleibt nicht bei der Antwort stehen. Die Plattform verbindet sich über APIs mit CRM, ERP, Ticket- und Policy-Systemen, legt Fälle an, fordert Dokumente an, ändert Stammdaten und schließt Vorgänge end-to-end ab, ohne dass ein Mitarbeitender eingreifen muss.

Die Integration läuft in der bestehenden Systemlandschaft: keine Migration, kein Systemwechsel, kein Vendor Lock-in. DSGVO-konform, gehostet in Deutschland.

octonomy

NÄCHSTER SCHRITT

Zeigen Sie uns Ihren Use Case.

In 30 Minuten analysieren wir gemeinsam, wo octonomy in Ihrem Servicebetrieb den größten Effekt erzielt. Konkret, ohne Verpflichtung.

Demo anfragen: octonomy.ai/demo ↗

octonomy AI GmbH, Hansaring 61, 50670 Köln, info@octonomy.ai



ÜBER OCTONOMY

Die Menschen hinter der Plattform

Die octonomy AI GmbH entwickelt KI-Lösungen für wissensintensive Serviceprozesse. Die Plattform verarbeitet komplexe Dokumentationen visuell, lässt sich in bestehende Systemlandschaften integrieren und automatisiert Serviceprozesse mit nachweislich hoher Antwortqualität.

Über octonomy

Die octonomy AI GmbH entwickelt KI-Lösungen für wissensintensive Serviceprozesse. Die Plattform verarbeitet komplexe Dokumentationen visuell, lässt sich in bestehende Systemlandschaften integrieren und automatisiert Serviceprozesse mit nachweislich hoher Antwortqualität.

Kontakt

Kontakt: info@octonomy.ai

Web: www.octonomy.ai

Redaktion

Maren Kaspers, VP Marketing, octonomy AI GmbH

Valeriia Levchuk, Marketing Manager, octonomy AI GmbH

LEKTORAT

Frank Piotraschke, Chief Revenue Officer, octonomy AI GmbH

Raj Khamkar, Team Lead Implementation, octonomy AI GmbH

Urheberrecht

© 2026 octonomy AI GmbH. Alle Rechte vorbehalten.

Alle Inhalte dieser Publikation, insbesondere Texte, Grafiken, Tabellen und Abbildungen, sind urheberrechtlich geschützt. Vervielfältigung, Verbreitung und öffentliche Wiedergabe, auch auszugsweise, sind ohne schriftliche Genehmigung der octonomy AI GmbH nicht gestattet. Zitate mit Quellenangabe sind ausdrücklich erlaubt.

Haftungsausschluss

Die in dieser Publikation enthaltenen Informationen wurden mit größtmöglicher Sorgfalt zusammengestellt. octonomy AI GmbH übernimmt keine Gewähr für die Vollständigkeit, Richtigkeit und Aktualität der Inhalte. Die Nutzung der Inhalte erfolgt auf eigene Gefahr.

Externe Links wurden zum Zeitpunkt der Verlinkung geprüft. Für den Inhalt verlinkter externer Seiten übernimmt octonomy AI GmbH keine Haftung.