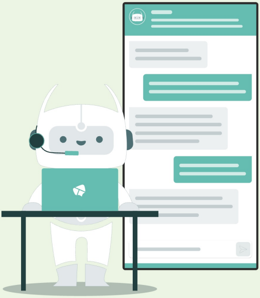
HERDT

Von generativer KI zu Agentic AI: Wie der HERDT-Verlag mit agentischer KI seinen Kundenservice nachhaltig entlastet hat

Interview mit Wilfried Fuchs • IT Cloud Engineer, HERDT-Verlag • Juni 2026

Case Study herunterladen

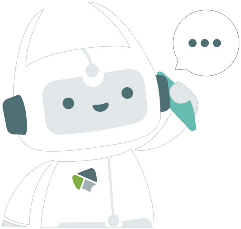


AUSGANGSPUNKT

Mehr als ein Chatbot: Wie der HERDT-Verlag mit KI in den Kundenservice gestartet ist

Der HERDT-Verlag für Bildungsmedien GmbH ist einer der führenden Verlage für IT-Lehr- und Lernmedien im deutschsprachigen Raum. Mit über 500 Themen rund um die digitale Arbeitswelt – von Microsoft 365 über Programmierung bis hin zu Netzwerken und KI beliefert er Unternehmen, Schulen und Weiterbildungseinrichtungen in Deutschland, Österreich und der Schweiz.Ihre Kunden stellen regelmäßig dieselben Fragen: „Wo ist meine Bestellung?“ “Welche Neuerscheinungen gibt es?” “Wie nutze ich einen Gutscheincode?” Diese Routineanfragen liefen über Mail, Telefon und Kontaktformular und belegten Kapazitäten im Kundenservice, die für komplexere Anliegen gebraucht wurden.

Die Entscheidung fiel auf KI im Kundenservice. Der erste Schritt war ein traditioneller generativer KI-Agent in Kombination mit NLU-Erkennungstechnologien: eine solide Lösung für den Einstieg, die Standardanfragen zuverlässig abdeckte. Doch im laufenden Betrieb wurden die Grenzen zunehmend sichtbar: Kunden formulierten frei, stellten mehrere Fragen auf einmal, und das System kam ins Straucheln. Was als guter Start funktionierte, reichte für den Alltag nicht mehr aus.



HERDT

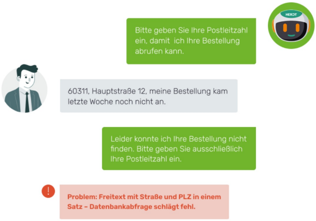
Wilfried Fuchs

IT Cloud Engineer • HERDT-Verlag für Bildungsmedien GmbH

DAS PROBLEM

Wo der generative KI-Agent an seine Grenzen stieß

GENERATIVER BERTI IM LIVEBETRIEB



Bitte geben Sie Ihre Postleitzahl ein, damit ich Ihre Bestellung abrufen kann.

60311, Hauptstraße 12, meine Bestellung kam letzte Woche noch nicht an.

Leider konnte ich Ihre Bestellung nicht finden. Bitte geben Sie ausschließlich Ihre Postleitzahl ein.

Problem: Freitext mit Straße und PLZ in einem Satz – Datenbankabfrage schlägt fehl.

Der generative KI-Agent Berti funktionierte zuverlässig, solange Kunden präzise formulierten und nur eine Frage auf einmal stellten. Beides war im Alltag selten der Fall. Freitext überforderte ihn häufig, kombinierte Anliegen wurden ignoriert und neue Anfragearten erforderten eine manuell gepflegte Erkennungsregel. Je breiter das Produktportfolio wurde, desto schneller wuchs das Regelwerk und desto mehr Lücken entstanden darin.

Freitext überforderte den generativen Agenten

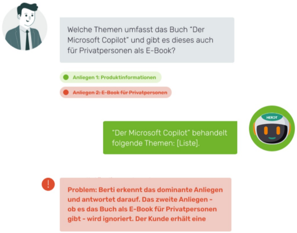
Generative KI-Agenten sind darauf ausgelegt, strukturierte Eingaben zu verarbeiten. Gibt ein Kunde mehr oder anderes an als erwartet, verliert der Agent den Faden. Ein konkretes Beispiel aus dem HERDT-Alltag: Berti fragte nach der Postleitzahl für eine Bestellabfrage. Kunden antworteten aber häufig mit mehr als erwartet: Postleitzahl und Straße in einem Satz oder gleich die halbe Lieferadresse. Der generative Agent konnte die relevante Information nicht herausfiltern, die Datenbankabfrage schlug fehl, die Antwort war falsch.

Eine Frage zu viel für den generativen Berti

Kombinierte Anliegen wurden ignoriert.

Fragte ein Kunde: „Welche Themen umfasst das Buch ‚Der Microsoft Copilot' und gibt es dieses auch für Privatpersonen als E-Book?", bearbeitete der traditionelle Berti nur einen Teil der Anfrage. Das zweite Anliegen, die E-Book-Verfügbarkeit für Privatpersonen, wurde übergangen.Der Grund liegt in der Orchestrierung: Der bisherige Berti arbeitete mit einer NLU, die pro Anfrage genau einen Intent erkennen konnte, das dominante Anliegen. Ein zweites gleichwertiges Anliegen in derselben Nachricht fiel dabei durch das Raster. In der Praxis bedeutete das: Der Kunde erhielt eine unvollständige Antwort.Dazu kam ein weiteres Problem: Der HERDT-Verlag hat ein breites Sortiment mit vielen ähnlichen Produkten und überschneidenden Prozessen. Der traditionelle generative KI-Agent arbeitete mit fest definierten Erkennungsregeln, die für jede Anfrageform manuell angelegt und gepflegt werden mussten. Mit wachsendem Portfolio und ähnlichen Anfragen kam das System regelmäßig durcheinander.

GENERATIVER BERTI IM LIVEBETRIEB



Welche Themen umfasst das Buch "Der Microsoft Copilot" und gibt es dieses auch für Privatpersonen als E-Book?

● Anliegen 1: Produktinformationen

● ~~Anliegen 2: E-Book für Privatpersonen~~

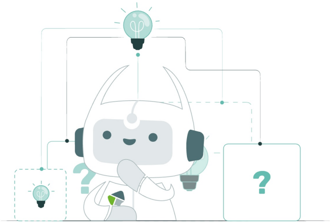
"Der Microsoft Copilot" behandelt folgende Themen: [Liste].

Problem: Berti erkennt das dominante Anliegen und antwortet darauf. Das zweite Anliegen - ob es das Buch als E-Book für Privatpersonen gibt - wird ignoriert. Der Kunde erhält eine

> „Für jede neue Anfrageart musste ich vorher eine eigene Erkennungsregel schreiben und pflegen. Je mehr Fälle dazukamen, desto unübersichtlicher wurde das System. Mit dem agentischen Berti beschreibe ich einfach, was ein Tool kann - und er weiß selbst, wann er es einsetzt."
>
> **Wilfried Fuchs**
>
> IT Cloud Engineer, HERDT-Verlag

DIE LÖSUNG

Die Lösung: ein agentischer KI-Agent



Der Wechsel von generativer zu agentischer KI war keine technische Notlösung, sondern eine bewusste Entscheidung für einen grundlegend anderen Ansatz.

Was ist Agentic AI und warum macht es den Unterschied?

Generative KI beantwortet eine Frage. Agentic AI löst ein Problem. Der Unterschied klingt klein, im Livebetrieb ist er entscheidend.Für den HERDT-Verlag war genau das der entscheidende Punkt. Ein KI-Agent, der nur dann funktioniert, wenn Kunden präzise formulieren, löst kein echtes Serviceproblem. Er verlagert es nur.

Generative AI vs. Agentic AI: Der direkte Vergleich

Kriterium	**Traditioneller Berti** Hybrider Bot – NLU + Generative KI	**Neuer Berti** Agentischer KI-Agent
ANFRAGEN VERSTEHEN	Präzise Eingabe nötig – Abweichungen führen zu Fehlern.	Versteht Freitext, erkennt Kontext, ordnet Informationen automatisch zu.
ANTWORT GENERIEREN	LLM beantwortet mit festem Prompt genau eine Anfrage – beschränkt auf hinterlegtes Wissen.	Kann Tools aufrufen – nicht auf Prompt-Wissen beschränkt.
MEHRFACHANLIEGEN	Pro Anfrage wird nur ein Intent erkannt – zweites Anliegen fällt durch das Raster.	Bearbeitet mehrere Anliegen in einer Nachricht, priorisiert eigenständig.
DATENBANKABFRAGEN	Starrer Abfragepfad – bereits genannte Infos werden ignoriert und müssen neu eingegeben werden.	Erkennt Eingabeparameter (PLZ, Bestellnr.) aus dem Gesprächsverlauf – fehlende gezielt nachgefragt.
KONFIGURATION	Manuelle Pflege von Erkennungsregeln für jede Anfrageart.	2–3 Wochen Einrichtung via Toolbeschreibungen (Teil des Prompts) – danach eigenständig.
PFLEGEAUFWAND / WOCHE	Ca. 3 Stunden (Wissen + Erkennungsregeln).	Ca. 30 Minuten (hochgerechnet aus ersten 2 Wochen).
NUTZERFÜHRUNG	Mechanisch und starr: fixer Pfad, bricht bei falscher Eingabe ab.	Kontextsensitiv: erkennt Gesagtes, hält Gesprächsfaden, fragt nur nach, was fehlt.

LIVE EINSATZ

Wo der generative KI-Agent an seine Grenzen stieß

Bestellstatus in Echtzeit: wie Berti jetzt antwortet

Was passiert, wenn ein Kunde schreibt: „Ich habe letzte Woche zehn Exemplare vom Excel-Grundkurs bestellt, drei davon fehlen in der Lieferung, Kundennummer 8821.“?

Der generative Berti kam damit nicht zurecht. Er arbeitete mit fest definierten Erkennungsmustern: Erst wenn der Kunde die Bestellnummer in genau der erwarteten Form eingab, landete die Anfrage im richtigen Ablauf. Wich die Formulierung zu stark ab, erkannte das System die Absicht gar nicht. Und selbst wenn es funktionierte, folgte ein starrer Pfad: Informationen, die der Kunde bereits genannt hatte, mussten trotzdem erneut eingegeben werden. Der Agent konnte nicht erkennen, was er schon wusste.

Der agentische Berti funktioniert grundlegend anders. Er weiß durch die Toolbeschreibung, welche Eingabeparameter für eine Bestellabfrage benötigt werden, zum Beispiel Kundennummer und Bestellmenge. Er erkennt eigenständig, ob diese Werte in der Nachricht bereits vorhanden sind, und geht direkt zur Abfrage über. Fehlt ein Wert, fragt er gezielt danach, präzise, nicht pauschal. Laut Wilfried Fuchs stimmt das Ergebnis nahezu zu 100 Prozent.

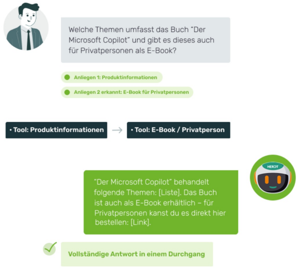
> „Kunden bekommen ihre Bestellinformationen jetzt direkt im Chat - und das Ergebnis stimmt, egal wie die Anfrage formuliert war."
>
> **Wilfried Fuchs**
>
> IT Cloud Engineer, HERDT-Verlag

Mehrere Fragen, eine Antwort: Wie der agentische Berti priorisiert

Einer der größten praktischen Vorteile der agentischen KI ist die Fähigkeit, mehrere Anliegen in einer Nachricht gleichzeitig zu verarbeiten. Der agentische Berti erkennt alle Anliegen, bewertet eigenständig die sinnvolle Reihenfolge und arbeitet sie mit Hilfe seiner bereitgestellten Tools vollständig ab, ohne dass der Kunde zweimal fragen muss.

Was hier im Hintergrund passiert, ist entscheidend: Der Agent ruft nicht einfach zweimal dieselbe Logik auf. Er erkennt, dass beide Anliegen unterschiedliche Prozesse erfordern, und startet sie gezielt. Für die Produktinformationen greift er auf den Wissens-Tool-Flow zurück. Für die E-Book-Frage startet er einen eigenen Flow und übergibt dabei automatisch den Eingabeparameter „Privatperson“ – weil er aus dem Freitext verstanden hat, für wen das Buch gedacht ist. Zwei unterschiedliche Prozesse, ein Durchgang, eine vollständige Antwort. Zwei Beispiele, ein gemeinsames Muster: Der agentische Berti versteht, was Kunden meinen, nicht nur, was sie tippen. Was das in der Einrichtung bedeutet, war für Wilfried Fuchs die nächste entscheidende Frage.

AGENTISCHER BERTI IM LIVEBETRIEB

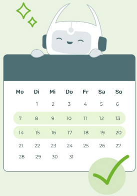


Welche Themen umfasst das Buch “Der Microsoft Copilot” und gibt es dieses auch für Privatpersonen als E-Book?

● Anliegen 1: Produktinformationen
● Anliegen 2 erkannt: E-Book für Privatpersonen
• Tool: Produktinformationen
→
• Tool: E-Book / Privatperson
HERDT

“Der Microsoft Copilot” behandelt folgende Themen: [Liste]. Das Buch ist auch als E-Book erhältlich – für Privatpersonen kanst du es direkt hier bestellen: [Link].

✓ Vollständige Antwort in einem Durchgang



Die Umstellung: 2–3 Wochen, kein Code, keine IT-Abteilung nötig

Die Umstellung auf den agentischen Berti war weniger eine technische als eine konzeptionelle Aufgabe: Statt Erkennungsregeln für jede Anfrageform zu pflegen, werden Toolbeschreibungen hinterlegt: kurze Texte, die erklären, was ein Tool kann und wann es sinnvoll eingesetzt wird. Diese Beschreibungen fließen direkt in den Prompt ein, mit dem das Sprachmodell angesprochen wird. Der Agent weiß dadurch, welche Tools ihm zur Verfügung stehen, und entscheidet eigenständig, welches er bei welcher Anfrage einsetzt. Die Wissensquellen bleiben dieselben, die Art, sie zu pflegen, wird einfacher. Wer die Logik einmal verstanden hat, kommt schnell rein.

Wilfried Fuchs brauchte 2–3 Wochenfür die Einrichtung. Seitdem läuft der agentische Berti eigenständig.

> „Das ist kein IT-Thema. Auch Kolleginnen und Kollegen ohne technischen Hintergrund können das konfigurieren. Ein logisches Grundverständnis genügt - mehr braucht es wirklich nicht."
>
> Wilfried Fuchs
>
> IT Cloud Engineer, HERDT-Verlag

Was das in der Praxis bedeutet: Die besten Konfiguratorinnen und Konfiguratoren für einen KI-Agenten sind nicht zwingend in der IT zu finden, sondern in den Teams, die das Wissen täglich leben. Marketing kennt die Produktinhalte, Kundenservice kennt die häufigsten Anfragen, Vertrieb kennt die Einwände. Sie wissen am besten, was ein KI-Agent wissen muss.

Was die Umstellung konkret gebracht hat, zeigen die Zahlen aus den ersten zwei Wochen nach Go-live.

ERGEBNISSE

Das Ergebnis: messbare Entlastung und deutlich weniger Pflegeaufwand

76%

5Std.

ANFRAGEN IN 2 WOCHEN
autmatisch beantwortet

ARBEITSZEIT GESPART
in den ersten 2 Wochen

>7/Tag

ANFRAGEN PRO TAG
tendenz steigend

24/7

VERFÜGBARKEIT
auch Sonn- und Feiertage

In den ersten zwei Wochen nach Go-live hat Berti 76 Kundenanfragen selbstständig beantwortet, Anfragen, die sonst direkt beim Kundenservice gelandet wären. Das entspricht mehr als 7 Anfragen pro Tag, Tendenz steigend. Pro Servicemitarbeiterin bedeutet das: weniger Routinelast, mehr Raum für die Anliegen, die wirklich Aufmerksamkeit brauchen. Was das in Arbeitszeit bedeutet: Bei jeder dieser Anfragen wäre früher ein direkter Kontakt mit dem Kunden nötig gewesen. Beim HERDT-Verlag dauerte eine solche Standardanfrage im Schnitt rund 4 Minuten. Auf die 76 automatisch beantworteten Anfragen hochgerechnet ergibt das eine eingesparte Bearbeitungszeit von rund 5 Stunden in den ersten zwei Wochen. Zum Vergleich: Branchenanalysen gehen im Call-Center-Bereich von durchschnittlich 6 Minuten pro Kontakt aus.Ein weiterer Vorteil, der schnell spürbar wird: Berti unterstützt Kunden auch außerhalb der Geschäftszeiten sowie an Sonn- und Feiertagen. Gerade im Weiterbildungsumfeld entstehen Fragen auch abends oder am Wochenende, genau dann ist Berti verfügbar.

> „Unser Kundenservice hat die Verbesserung nicht durch Auswertungen wahrgenommen, sondern ganz direkt: Das Telefon klingelte seltener, weniger E-Mails kamen rein. Wenn die agentische KI ihre Arbeit richtig macht, kriegt der Kundenservice davon gar nichts mit. Das ist das Schöne daran."
>
> **Wilfried Fuchs**
> IT Cloud Engineer, HERDT-Verlag

Das Ergebnis: messbare Entlastung und deutlich weniger Pflegeaufwand

3Std.

PFLEGEAUFWAND ALT
pro Woche (generativ)

30Min.

PFLEGEAUFWAND NEU
pro Woche (agentisch, hochger.)

~80%

WENIGER AUFWAND
rechnerisch

Die generative Version von Berti erforderte regelmäßig rund 3 Stunden Pflege pro Woche: eine Stunde für die Aktualisierung von Wissensinhalten wie neue Produkte oder geänderte Lieferzeiten, zwei Stunden für die Anpassung der Erkennungsregeln. Denn im Livebetrieb stellte sich immer wieder heraus, dass Kunden andere Formulierungen nutzten als intern erwartet.

Mit der agentischen Version hat sich das grundlegend verändert. In den ersten zwei Wochen nach Go-live hat Wilfried Fuchs lediglich widersprüchliche Formulierungen in drei Wissenselementen korrigiert (ca. 45 Minuten) und zwei neue Produkte in die Wissensbasis aufgenommen (ca. 10 Minuten). Zusammen: 55 Minuten in zwei Wochen – hochgerechnet rund 30 Minuten pro Woche.

> „Nach zwei bis drei Wochen Einrichtungsaufwand läuft Berti jetzt eigenständig. Ich behalte die Kundenantworten im Blick und schaue, ob die Leute zufrieden sind. Technische Wartung? Kaum noch ein Thema. Es geht nur noch darum, das hinterlegte Wissen aktuell zu halten."
>
> **Wilfried Fuchs**
>
> IT Cloud Engineer, HERDT-Verlag

ZUSAMMENARBEIT

Die Zusammenarbeit: Betreuung auf Augenhöhe

Einführung mit Begleitung: Wie DialogBits den Unterschied macht

Ein Faktor, den Wilfried Fuchs besonders hervorhebt: die Begleitung durch das DialogBits-Team während der Einführung. Die Zusammenarbeit mit Anna, der Customer-Success-Managerin von DialogBits, verlief in sechs gemeinsamen Meetings über Google Meet, vom ersten Orientierungsgespräch bis zum Go-Live.

Am Anfang stand kein Tool, sondern Kontext: Anna hielt Wilfried Fuchs über die Entwicklungen rund um Agentic AI auf dem Laufenden, bevor überhaupt konfiguriert wurde. Als es so weit war, kopierte sie den bestehenden Berti und richtete ihn so ein, dass Wilfried anhand seines eigenen Use-Cases sehen konnte, wie Agentic AI in der Praxis funktioniert. Von da an übernahm er selbst: Er schrieb die Toolbeschreibungen, testete den neuen Berti und stellte ihn dem alten direkt gegenüber, der zu diesem Zeitpunkt noch live war.In den Folge-Meetings zeigte Wilfried, was er gebaut hatte. Anna gab Feedback, beantwortete Fragen und war bei dringenden Themen auch zwischen den Terminen per Mail erreichbar. Anna begleitete den Prozess mit sechs gemeinsamen Meetings – vom ersten Onboarding bis zum Go-Live.

Case Study herunterladen

Kundenkommunikation auf Autopilot: Erlebe DialogBits live!

Jetzt Demo sichern & 14 Tage gratis testen – ganz ohne Verpflichtung.

Demotermin buchen

„Die Betreuung durch Anna von DialogBits war der Grund, warum wir immer das Gefühl hatten: Da ist jemand, der wirklich mitdenkt."

Wilfried Fuchs

IT Cloud Engineer, HERDT-Verlag