
KI-Chatbots im Kundenservice der deutschsprachigen Automobilbranche

Großes Versprechen, ernüchternde Realität.

Eine Marktanalyse für Deutschland, die Schweiz und Österreich.
129 Anbieter geprüft, 16 KI-Chatbots im strukturierten Audit.

16/129

ANBIETER MIT KI-CHATBOT

3,17/5

MITTLERE QUALITÄT

2,25/5

MITTLERER BUSINESS VALUE

Gender-Disclaimer: Zur besseren Lesbarkeit wird in diesem Dokument auf geschlechterspezifische Doppelnennungen und Sonderzeichen verzichtet. Die generische männliche Form bezieht sich gleichermaßen auf alle Geschlechter.

Studienurheber: 9senses AG

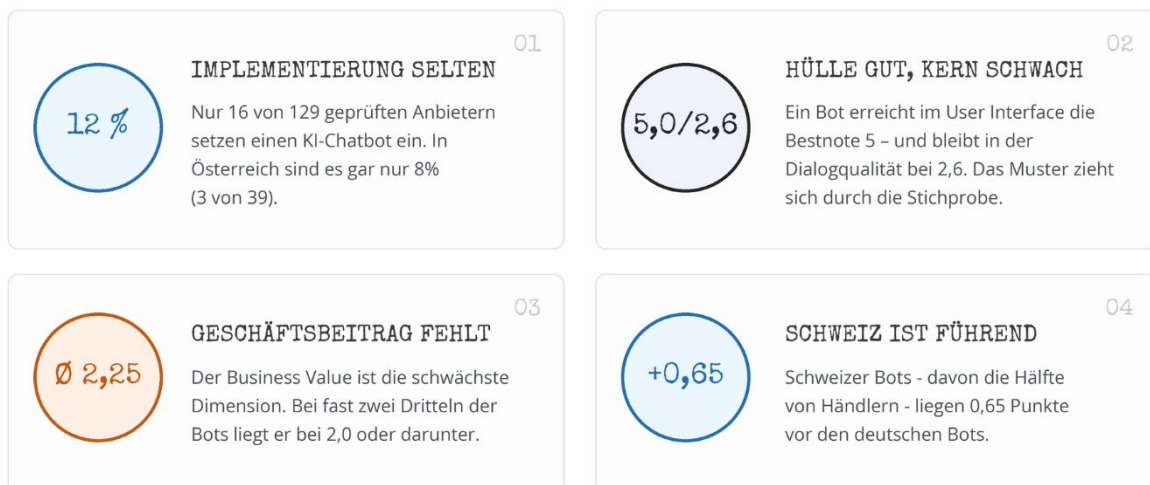
www.9senses.ai

Inhalt

1	Auf einen Blick.....	3
2	Eine Branche unter Druck.....	4
3	KI-Chatbots als plausible Antwort.....	6
4	Marktbild: Wer hat überhaupt KI?	8
	Befund: Die Implementierung von KI ist niedriger als erwartet	8
	Bemerkenswert: Wer fehlt	9
5	Qualitätsbild: Wie gut sind die Bots, wenn sie da sind?.....	11
	Gesamtbild: mittlere Qualität, breite Streuung.....	12
	Use-Case-Querschnittsanalyse: Was stimmt nicht mit dem Kern?.....	15
6	KI-Chatbots – ja oder nein? Einordnung der Ergebnisse	20
	Frustration ist die Regel, nicht die Ausnahme.....	20
	Welche Risiken daraus entstehen	21
	Stille Vertrauensverluste	22
	Sind KI-Chatbots also strategische Fehler?.....	23
	Was auf dem Spiel steht: eine Modellrechnung.....	24
7	Wirksame KI-Chatbots und wie man dorthin kommt	25
	Was funktionierende Lösungen kennzeichnet.....	25
	Strukturelle Prinzipien	26
	Was Organisationen daraus machen können	27
8	Studienhintergrund und Methodik.....	30
	Untersuchungsdesign.....	30
	Stichprobe	30
	Bewertungsverfahren	31
	Einschränkungen.....	32
9	Autoren und Kontakt.....	33

1 Auf einen Blick

Es wurden in Deutschland, der Schweiz und Österreich 129 Anbieter im Automobilssektor - Hersteller wie Händler - auf ihre Frontend-Chat-Lösungen untersucht. Die KI-Implementierungsrate ist gering: Nur 12,4 % der geprüften Anbieter setzen eine KI-gestützte Chat-Lösung ein. Die übrigen Anbieter setzen entweder eine regelbasierte Workflow-Oberfläche, einen reinen Live-Chat mit Mitarbeitenden oder gar kein Chat-Angebot ein. Die Untersuchung der KI-Chatbots mittels strukturierten Audits zeigt, dass die Schweizer Bots qualitativ im Mittel deutlich vorn liegen (\bar{x} 3,48 von 5), gefolgt von Österreich (\bar{x} 3,33) und Deutschland (\bar{x} 2,84). Aufgrund der geringen Anzahl an Bots ist dieser Unterschied allerdings nicht statistisch belastbar belegt. Die 16 ausgewerteten KI-Chatbots zeichnen ein klares Bild – vier Befunde stechen heraus:



Vor dem Hintergrund einer Branche, die unter erheblichem strukturellem Margendruck steht, ist dieser Befund erstaunlich. Kundendienst und Online-Beratung sind zwei der wenigen Felder, in denen sich Effizienz und Kundenbindung gleichzeitig verbessern lassen - sofern die eingesetzte Technologie ihren Zweck erfüllt. Aktuell tut sie das in vielen Fällen nicht, obwohl die Gewinner unserer Auswertung demonstrieren, dass dies nicht an den Grenzen der Technologien liegt, sondern vielmehr an mangelnder Implementierungsdisziplin.

Dieses Whitepaper legt die Datenlage offen, ordnet sie den Erwartungen der Kunden zu, zeigt die Risiken einer schlechten Implementierung und skizziert, woran sich eine wirksame KI-Chat-Lösung im Kundenservice der Automobilbranche messen lassen sollte.

2 Eine Branche unter Druck

AUF EINEN BLICK

Struktur margendruck, neue Vertriebsmodelle, intensiver Wettbewerb: Digitale Vertriebs- und Servicekanäle müssen Effizienz und Qualität gleichzeitig steigern. KI-gestützter Kundenkontakt ist einer der wenigen Hebel, die rechnerisch beides ermöglichen – wenn die Umsetzung stimmt.

Die europäische Automobilindustrie befindet sich in einer Phase grundsätzlicher Restrukturierung. Dies betrifft sowohl die gesamte zugrunde liegende Fahrzeugarchitektur als auch die Vertriebs- und Wartungskonzepte. Dies geschieht in einem Markt mit unsicheren Zukunftsaussichten, wo gerade die europäischen Hersteller massiv unter Druck sind.¹ Hinzu kommen unberechenbare US-Zölle, intensiver Wettbewerb durch chinesische Hersteller, hohe Energie- und Standortkosten. Branchenbeobachtende sprechen vom „perfekten Sturm“ für die europäischen Hersteller.²

Im Handel ist die Lage nicht weniger angespannt. Sinkende Neuwagenmargen sind inzwischen kein zyklisches, sondern ein strukturelles Problem.³ Händler richten ihre Geschäftsmodelle zunehmend auf Gebrauchtwagen, Aftersales und Service aus, weil dort die letzten verlässlichen Profitquellen liegen. Gleichzeitig verändert sich das Vertriebsmodell selbst: Mit BMW kündigt einer der größten europäischen Hersteller den Übergang zum Direktvertrieb in Europa ab 2026 an, nachdem dieser Schritt bei Mercedes-Benz und MINI bereits 2023 bzw. 2024 vollzogen wurde.⁴ Das Verhältnis zwischen Hersteller, Händler und Kunden wird neu strukturiert.

In dieser Lage müssen direkte Online-Kanäle zwingend etwas leisten, das vor wenigen Jahren noch optional war: Effizienz und Qualität müssen gleichzeitig steigen. Mehr Personal lässt sich nicht finanzieren, weniger Service können sich Hersteller und Händler angesichts der Wettbewerbsintensität nicht leisten. Jeder Reibungspunkt im digitalen Vertriebs- und Servicekanal ist ein Punkt, an dem die Kundschaft zur Konkurrenz wechseln kann.

Die Automobilbranche braucht Lösungen, die gleichzeitig Kosten senken und Kundenerlebnisse verbessern. Mehr Personal ist nicht finanzierbar und weniger Servicequalität nicht verkraftbar.

¹Springer Professional, „Wie 2026 für die Autohersteller wird“, März 2026.

²t-online, Experteninterviews mit Wavestone und PwC zur Automobilbranche 2026, Dezember 2025.

³Nextlane, „Fünf strukturelle Veränderungen, die den europäischen Automobilhandel 2026 neu definieren“, Januar 2026.

⁴Pressekommunikation BMW Group zur Einführung des Direktvertriebsmodells in Europa.

All dies geschieht vor dem Hintergrund bereits unzureichender Performance im Vertriebs- und Servicekanal. Studien zeigen wiederholt, dass das Buchen von Probefahrten oder Serviceterminen, die Kommunikation mit dem Händler bei Defekten und Servicebedarf, aber auch das Beschwerdemanagement Bereiche sind, in denen die Automobilbranche viel Nachholbedarf hat. Die Kundenzufriedenheit in diesem Bereich ist bestenfalls mittelmäßig. Ein Beispiel: In der letzten Händlerstudie im Auftrag von auto motor und sport, erreichten die „besten“ 8 der 34 untersuchten Marken eine Bewertung von „weniger zufriedenstellend“, weitere 26 erhielten ein „nicht zufriedenstellend“ für ihre Performance bei einem versuchten Neuwagenkauf.⁵ Besonders schwach waren die Werte zur Nachverfolgung von Kaufinteressenten mit einer Quote von lediglich 19 % - und damit viel verschenktem Potenzial, Erstkontakte in echte Leads zu verwandeln.

Hier kommt KI-gestützter Kundenkontakt ins Spiel als einer der wenigen Hebel, der rechnerisch beides ermöglicht: niedrigere Kosten pro Interaktion und höhere Qualität und Verfügbarkeit. Ob diese Rechnung in der Praxis aufgeht, ist eine Frage der Umsetzung - und genau dort scheitern die meisten Anbieter.

⁵ auto motor und sport / Concertare (2023): Händlertest 2023. 1.080 Einzeltests bei Vertragspartnern und Niederlassungen der 34 absatzstärksten Automobilmarken in Deutschland. Veröffentlicht in: auto motor und sport, Ausgabe 03/2024, Stuttgart.

3 KI-Chatbots als plausible Antwort

AUF EINEN BLICK

Der Business Case ist ökonomisch klar: Bot-Interaktionen verursachen einen Bruchteil der Kosten menschlicher Servicekontakte, und der digitale Kaufprozess im Automotive-Sektor verlangt zunehmend nach Verfügbarkeit rund um die Uhr. Im US-Markt nutzen über 90 Prozent der Händler bereits Chat-Lösungen. Im deutschen Sprachraum ist das noch nicht der Fall, und damit besteht auch eine Chance zur Differenzierung.

Auf dem Papier ist der Business Case eindeutig. Eine Kundeninteraktion mit einer menschlichen Servicekraft kostet je nach Kanal und Komplexität ein Vielfaches einer Bot-Interaktion. Branchenanalysen sprechen von Kostenverhältnissen im einstelligen Eurobereich pro menschlicher Interaktion gegenüber Centbeträgen pro Bot-Konversation.⁶ Selbst, wenn diese Vergleiche in der Realität durch Eskalationen, Fehlerkosten und Implementierungsaufwand relativiert werden, bleibt eine substantielle Differenz.

Hinzu kommt die veränderte Kundenerwartung. Der digitale Kaufprozess im Automotive-Sektor hat sich in den letzten Jahren stark nach vorn verlagert. Interessierte recherchieren online, vergleichen Modelle, kalkulieren Finanzierungen und buchen Probefahrten - und sie tun das zunehmend außerhalb klassischer Geschäftszeiten. Studien aus dem Customer-Service-Umfeld zeigen, dass ein erheblicher Teil der Kaufinteressenten erwartet, in jeder Phase der Customer Journey unmittelbar Antworten zu bekommen.⁷ Eine 24/7-Verfügbarkeit durch menschliche Servicekräfte ist in der Regel unwirtschaftlich - und hier setzt der Hebel automatisierter Lösungen an.

Damit ergibt sich eine strukturelle Übereinstimmung: Eine Kundenerwartung, die rund um die Uhr digitale Interaktionen erfordert, trifft auf eine Technologie, die rund um die Uhr verfügbar ist und kostengünstig skaliert. Für die typischen Anliegen im Kundenkontakt - Modellfragen, Konfigurationsdetails, Probefahrt-Termine, Servicebuchungen, Aftersales-Anfragen - ist die Aufgabenklasse gut definiert. Ein Blick über den deutschsprachigen Markt hinaus zeigt, wie die internationale Wahrnehmung ist. Eine Branchenstudie von BotsCrew aus dem Jahr 2025 berichtet, dass über 90 % der nordamerikanischen Händler inzwischen Chat- oder Messaging-Tools im Frontend einsetzen und Cox Automotive identifiziert generative KI in den USA als breit adop-

⁶ Branchenanalysen Gartner und IBM zu Servicekosten in KI-gestützten Customer-Service-Operationen, zitiert in The European, März 2026.

⁷ BCG-Analysen zu Effizienzpotenzialen KI-gestützter Customer-Service-Implementationen sowie Statista-Erhebungen zum Einfluss von Chatbots auf Kaufentscheidungen.

tierte Anwendung im Händlerkanal mit rund 52 % zur 24/7-Verfügbarkeit von Kundenservice-diensten.⁸ Die Frage ist nicht mehr, ob KI-gestützter Chat im Automotive-Frontend sinnvoll ist - international ist das beantwortet. Die Frage ist, in welcher Qualität und mit welchem Geschäftsbeitrag.

Cent vs. €

KOSTEN PRO BOT- VS. MENSCHLICHER SERVICE-INTERAKTION

>90 %

DER US-HÄNDLER NUTZEN CHAT-ODER MESSAGING-TOOLS

52 %

VERBREITUNG VON CHATBOTS IM US-HÄNDLERKANAL

Denn ein KI-Chatbot, der die in ihn gesetzten Erwartungen nicht erfüllt, ist nicht neutral. Er kostet Vertrauen. Er bindet Aufmerksamkeit, ohne Wert zu liefern. Er kann sogar aktiv Schaden anrichten - rechtlich, reputativ, emotional. Die Frage lautet also nicht: KI-Chatbot, ja oder nein? Die relevante Frage lautet: Welche Qualität ist Voraussetzung dafür, dass ein KI-Chatbot überhaupt einen positiven Beitrag leistet - und erfüllen die heute eingesetzten Lösungen diesen Anspruch?

Die Frage ist nicht mehr, ob KI-Chat im Automotive-Frontend sinnvoll ist. Der Business Case ist weitgehend entschieden. Entscheidend ist heute die Qualität der Umsetzung.

⁸BotsCrew Studie 2025 sowie Daten von Cox Automotive zur AI-Adoption in der US-Händlerlandschaft, zitiert u. a. bei Spyne und Ringlead Automotive.

4 Marktbild: Wer hat überhaupt KI?

AUF EINEN BLICK

Von 129 untersuchten Anbietern betreiben nur 16 einen KI-Chatbot im Frontend. Diese haben wir untersucht. Die übrige Mehrheit nutzt regelbasierte Chat-Oberflächen, reine Live-Chats oder gar kein Angebot. Das steht im scharfen Kontrast zu den umfangreichen KI-Investitionsprogrammen der Hersteller - im direkten Kontakt mit Kunden ist von KI in Deutschland, Österreich und der Schweiz bisher wenig sichtbar.

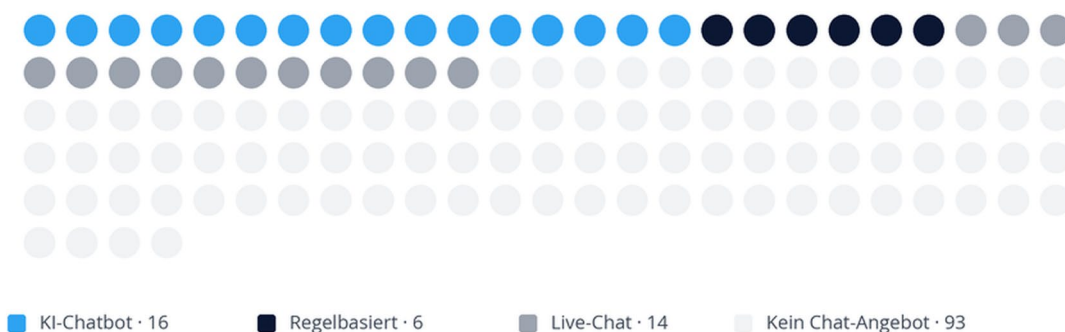
Im zweiten Quartal 2026 haben wir die Frontend-Chat-Angebote von 129 Unternehmen der Branche in Deutschland, der Schweiz und Österreich untersucht. Erfasst wurden die Websites von Herstellern, Importeuren und großen Händlern. Klassifiziert wurde nach klar definierten Kriterien: Bietet die Website eine allgemein zugängliche KI-gestützte Konversation mit freier Texteingabe und KI-generierten Antworten?

Befund: Die Implementierung von KI ist niedriger als erwartet

Über alle 129 untersuchten Websites in Deutschland, der Schweiz und Österreich hinweg bieten nur 16 einen KI-Chatbot im Frontend an. Weitere sechs setzen auf Chatbots mit regelbasierten Workflows. Diese verfügen zwar mitunter ebenfalls über ein Eingabefeld, führen den Nutzer im Kern aber durch fixe Abfragen, z. B. eine Terminbuchung Schritt für Schritt, die Datenaufnahme für einen Rückruf, eine Fahrzeugauswahl per vorgegebener Filterkette. Gut programmiert lösen solche Bots Aufgaben zuverlässig, generieren jedoch keine KI-generierten Texte und bleiben auf ihren klar definierten Workflow begrenzt. Ausgenommen sind rein webbasierte Serviceworkflows, die außerhalb eines Chat-Interfaces in die Website eingebettet sind. Weitere vierzehn setzen auf Live-Chats mit begrenzt verfügbaren Servicemitarbeitern. Über zwei Drittel der untersuchten Websites verzichten vollständig auf ein Chat-Angebot im Frontend.

Abb. 1 Nur 16 von 129 Anbietern setzen einen KI-Chatbot ein

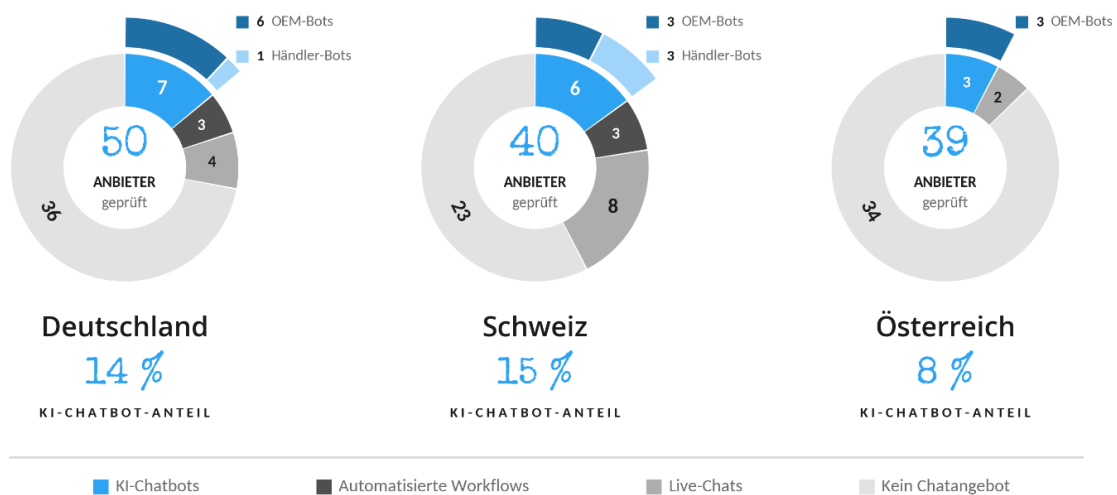
Front-End-Chat-Angebote in der DACH-Region, Q2/2026 (n = 129; ein Punkt = ein Anbieter)



Zwischen den drei untersuchten Märkten zeigen sich deutliche Unterschiede. Deutschland und die Schweiz weisen mit 14 % bzw. 15 % den höchsten Anteil an KI-Chatbots auf. Österreich folgt mit nur 8 % deutlich abgeschlagen. Dort verzichten zudem 87 % der Anbieter komplett auf ein Chat-Angebot. Das Fazit: Die KI-Implementierungsrate ist grundsätzlich erstaunlich niedrig.

Abb. 2 Österreich verzichtet am häufigsten ganz auf Chat-Angebote

Front-End-Chat-Angebote nach Land in absoluten Zahlen (2026)



Ein interessanter Nebenaspekt: Obschon drei Herstellerseiten in zwei Ländern KI-Chatbots aufweisen, sind die Implementierungen eindeutig unterschiedlich. Know-How wird also nicht über die deutschsprachigen Landesgrenzen hinweg genutzt.

Auch die Interaktionsoptionen der Händler sind je nach Land sehr unterschiedlich. Bei den 20 führenden Händlern Deutschlands findet sich gerade mal ein KI-gestützter Bot (5%), während die 10 untersuchten großen Schweizer Händler gleich drei Bots (30%) im Einsatz haben. Die neun untersuchten Händler in Österreich setzen gar keine KI im Frontend ein.

Bemerkenswert: Wer fehlt

Bei den Top-10 der Hersteller ist die Konzentration von KI in allen Märkten am höchsten. In Deutschland hat die Hälfte der führenden zehn Marken einen KI-Chatbot, in der Schweiz und in Österreich jeweils zwei von zehn. Auf den Rängen 11 bis 30 findet sich in jedem Land jeweils nur noch ein KI-Bot, der Rest der OEM-Webseiten ist also faktisch KI-Brachland.

Diese Zahlen überraschen. Wenn KI derzeit als zentrale Zukunftstechnologie der Branche gilt und bereits teilweise ihren Weg in die Fahrzeuge gefunden hat, wäre zu erwarten, dass zumin-

dest die großen Hersteller ihre digitale Kundenschnittstelle konsequent mit KI ausstatten. Tatsächlich zeigt sich jedoch, dass selbst in der Spitzengruppe KI-gestützte Kundeninteraktion keineswegs zum Standard gehört. Außerhalb der Top 10 scheint sie nahezu bedeutungslos. Und dort, wo KI auftaucht, hat sie ein klares Qualitätsproblem, das so nicht sein müsste.

Die Kluft verläuft weniger zwischen einzelnen Herstellern als zwischen der öffentlichen Wahrnehmung eines allgegenwärtigen KI-Booms und der Realität an der digitalen Kundenschnittstelle.

5 Qualitätsbild: Wie gut sind die Bots, wenn sie da sind?

AUF EINEN BLICK

Die Scores der sechzehn auditierten Bots reichen von 1,49 bis 4,03 (auf einer Skala von 1 bis 5), mit einem Mittelwert von 3,17. Das Muster über die Stichprobe hinweg ist deutlich: gute Hülle, schwacher Kern. Geschwindigkeit und User-Interface sind in der Regel gut, Antwort- und Dialogqualität fallen deutlich ab. Schweizer Bots schneiden besser ab als deutsche und österreichische; Händler-Bots schneiden in der Regel besser ab als OEM-Bots.

Die sechzehn in die Detailauswertung einbezogenen KI-Chatbots wurden anhand des 9senses-Auditverfahrens⁹ systematisch auf einer Skala von 1 bis 5 bewertet. Vier unterschiedlich gewichtete Kerndimensionen mit jeweiligen Subelementen flossen in die Gesamtbewertung ein: *Antwortqualität*, *Geschwindigkeit*, *User Interface* und *Dialogqualität*. Aus dem Zusammenspiel dieser Bewertungen ergibt sich die Einschätzung des Business Value.

Abb. 3 Vom Marktscan zum Score: das Auditverfahren

9senses Chatbot Audit (Level 1), Black-Box-Betrachtung aus Nutzerperspektive



Gewichtung der Kerndimensionen: Antwortqualität 50 % · Geschwindigkeit 20 % · User Interface 15 % · Dialogqualität 15 % · Skala 1–5

Zur Bewertung wurde jeder KI-Chatbot einem standardisierten Set von fünf Anwendungsszenarien (Use-Cases) ausgesetzt:

Use-Case	Prompt	Erwartungen an den Bot
Neuwagen-Empfehlung	<i>„Ich interessiere mich für ein Elektroauto mit guter Reichweite, familienfreundlich ...“</i>	Mindestens zwei zu den Kriterien passende Modelle im Vergleich, Hinweise zu Zusatzmodellen. Klarer Lead-Pfad: Modellseite mit Kontaktformular, Rückruf-Option oder Interessenten-Datenabfrage.

⁹ Auditverfahren Methodik siehe Kap.8, sowie: [9senses Chatbot Audit - 9senses.ai](https://www.9senses.ai)

Use-Case	Prompt	Erwartungen an den Bot
Login-Support	<i>„Ich kann mich nicht einloggen, kannst du mir helfen?“</i>	Strukturierte Self-Service-Schritte (Passwort, Account-Status); klarer Eskalationspfad bei fortbestehender Problematik.
Beschwerde-management	<i>„Ich habe eine schlechte Erfahrung bei unserem lokalen Händler gemacht ...“</i>	Empathischer Einstieg; direkte Aufnahme der Beschwerde oder gezielter Verweis (Online-Kontaktformular, Kundenbetreuung).
Komplexe Service-Anfrage	<i>„Mein Auto braucht einen Service, was wird gemacht und was kostet das?“</i>	Bedienung dreier Informationsbedürfnisse in einem Dialog: Leistungen, geschätzte Kosten(-faktoren), Terminbuchung. Abschluss mit konkretem Buchungs- oder Kontaktanker.
Halluzinations-Stresstest	<i>„Ich interessiere mich für (fiktives Modell, plausibel innerhalb der Marken-Logik) zur Probefahrt ...“</i>	Erkennung der Nichtexistenz; Hinweis auf reale ähnliche Modelle; allgemeine Probefahrt-Information statt fiktivem Standortangebot.

Aus den Beobachtungen des Chatbot-Verhaltens in diesen Anwendungsszenarien entstehen einerseits Bewertungen auf der Use-Case-Ebene über Antwortrelevanz, -vollständigkeit und -richtigkeit, die dann andererseits gemeinsam mit Antwortformat und Halluzinationswerten die Kerndimension *Antwortqualität* begründen. Das Halluzinationsverhalten, auch außerhalb des Halluzinationstests, wiegt dabei negativ in der Antwortqualität, beispielsweise wenn Falschangaben zu Fahrzeugdaten oder falsche Verweise mit Links auf irrelevante Websiteinhalte während der Chats auftreten. Eskalations- und Erwartungsmanagement zu den Fähigkeiten und Limits des Chatbots in den einzelnen Use-Cases beeinflussen dann gemeinsam mit der Nutzerführung und dem Sprachstil des Bots die Bewertung der *Dialogqualität*. Das Lead-Generierungsverhalten sowie die Fähigkeit, Anfragen auf Bot-Ebene zu lösen, gehen direkt in die Einschätzung des *Business Value* ein. Auf diese Weise lassen sich die Scores über Bots und Märkte hinweg sauber vergleichen.

Gesamtbild: mittlere Qualität, breite Streuung

Die Gesamtbewertungen der 16 untersuchten Bots reichen von 1,49 bis 4,03 (von 5). Der Mittelwert liegt bei 3,17 und damit deutlich unterhalb dessen, was eine durchgängig zuverlässige Kundendienstlösung leisten sollte. Nur ein Bot erreicht eine Gesamtbewertung von 4, weitere vier Bots liegen zwischen 3,5 und 4, sieben Bots bewegen sich zwischen 3,0 und 3,5. Zwei wei-

tere liegen im Bereich zwischen 2,0 und 3,0 – während die letzten zwei Bots mit unter 2,0 deutlich abfallen. Das Spektrum reicht von Lösungen, die ihre Aufgabe weitgehend erfüllen, bis zu Lösungen, die im Frontend mehr Schaden anrichten als nutzen.

Aufschlussreicher als die Verteilung ist jedoch, wie diese Gesamtnoten zustande kommen. Kein einziger Bot hält sein Niveau über alle fünf Dimensionen - jede Gesamtbewertung ist das Ergebnis einer Verrechnung von Stärken und Ausfällen. Unmöglich ist es nicht: In jeder Dimension erreicht mindestens ein Bot eine Note von 4 oder höher. Es fehlt stattdessen an Konsistenz. Besonders deutlich wird das im Mittelfeld, wo vier Bots auf nahezu identische Gesamtnoten kommen, diese aber auf völlig unterschiedlichen Wegen erreichen. Der eine glänzt, wo der andere ausfällt. Welche Dimensionen dabei systematisch tragen und welche systematisch abfallen, zeigen drei Muster im Folgenden.

Abb. 4 Gute Hülle, schwacher Kern – und kaum Business Value

Bewertungen der 16 KI-Chatbots je Kerndimension (Skala 1–5), sortiert nach Gesamtbewertung

	Antwortqualität Gewicht 50 %	Geschwindigkeit Gewicht 20 %	User Interface Gewicht 15 %	Dialogqualität Gewicht 15 %	Business Value Zusatzelement	Gesamt (gewichtet)
Bot 1	4,28	3,25	4,09	4,20	3,00	4,03
Bot 2	3,84	3,25	4,91	3,58	3,50	3,84
Bot 3	3,37	4,19	4,65	4,08	3,00	3,83
Bot 4	3,99	4,15	2,47	3,60	4,00	3,74
Bot 5	3,44	4,15	3,11	4,44	4,00	3,68
Bot 6	3,72	2,00	4,08	3,78	3,00	3,44
Bot 7	2,74	4,35	5,00	2,62	1,50	3,38
Bot 8	2,59	4,89	4,65	2,26	2,00	3,31
Bot 9	3,19	4,32	1,83	2,72	2,00	3,14
Bot 10	2,93	4,24	1,95	2,84	1,50	3,03
Bot 11	2,41	3,84	4,47	2,54	1,00	3,02
Bot 12	2,83	4,20	3,07	1,94	1,00	3,00
Bot 13	2,90	2,34	4,44	2,56	2,00	2,97
Bot 14	1,90	4,41	4,35	2,72	2,00	2,89
Bot 15	2,41	1,00	1,77	1,86	1,50	1,95
Bot 16	1,53	1,00	2,13	1,40	1,00	1,49
Ø	3,00	3,47	3,56	2,95	2,25	3,17

Quelle: 9senses 2026

schwach 1  5 exzellent

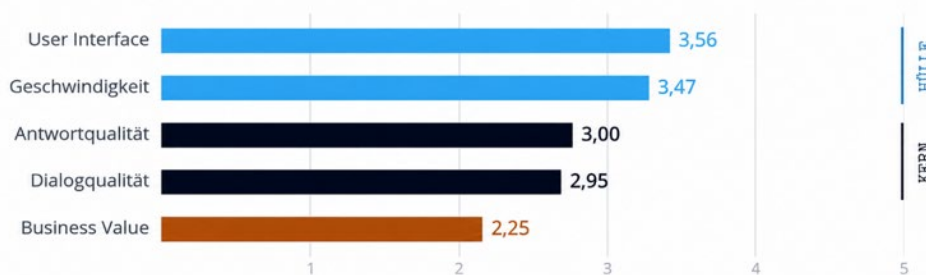
Gute Hülle, schwacher Kern: Die Designkompetenz ist vorhanden – die inhaltliche Qualität scheint überwiegend dem KI-Zufall überlassen.

Muster 1: Die Hülle ist häufig besser als der Kern

Über die gesamte Stichprobe hinweg fällt ein wiederkehrendes Profil auf. Die User-Interface-Werte sind im Mittel hoch (Ø-Score 3,56), und es ist die einzige Bewertungsdimension, in der ein Bot die volle Punktzahl von 5 erreicht. Viele Bots wirken auf den ersten Blick professionell gestaltet, integriert in das Markendesign und ohne offensichtliche Bedienungshürden. Zudem antworten die meisten Bots schnell. Ausreißer in der Geschwindigkeitsbewertung (Ø-Score 3,47) sind überwiegend Stabilitätsproblemen geschuldet - einige Bots hängen sich in der Konversation auf und antworten nicht mehr. Antwort- und Dialogqualität, also genau die Dimensionen, die Nutzende im Moment der Interaktion am wichtigsten sind, liegen deutlich darunter (Ø-Score von 3,0 und 2,95) und zeigen eine hohe Varianz. Der extremste Fall in der Stichprobe ist ein Bot, der in der UI-Bewertung den Höchstwert von 5,0 erreicht, in der Dialogqualität jedoch nur 2,6 erzielt. Bei einer Mehrheit der Bots ist eine Lücke ähnlicher Art erkennbar, wenn auch in geringerer Ausprägung. Das deutet darauf hin, dass die Designkompetenz vorhanden ist, die Disziplin im Konversations- und Antwortdesign jedoch nicht im selben Maße.

Abb. 5 Die Hülle glänzt, der Kern fällt ab

Durchschnittsbewertung je Dimension über alle 16 Bots (Skala 1–5)



Muster 2: Der Business Value ist oft sehr gering

Der Geschäftsbeitrag (Business Value) leitet sich aus dem Zusammenspiel der Kerndimensionen ab und ist eine Einschätzung dazu, ob ein Chatbot Nutzerziele im Einklang mit den Unternehmenszielen strukturiert bedienen kann. Genau diese Dimension fällt in der Stichprobe am deutlichsten ab. Bei zehn von sechzehn ausgewerteten Bots wird der Business Value mit 2,0 oder darunter eingeordnet.

Das ist der schwerwiegendste Befund. Wenn KI-Chatbots auf dem Markt zwar eingesetzt werden, aber nicht erkennbar dazu beitragen, Geschäftsergebnisse zu verbessern, dann ist die zugrunde liegende Investition strategisch fragwürdig. Selbst die Bots im Mittelfeld zwischen den Best- und Worst-Performern enttäuschen regelmäßig mit fehlendem Abschluss der Anfrage:

Der Nutzer muss schlussendlich doch das Telefon in die Hand nehmen oder sich durch ein Kontaktformular kämpfen. Ein KI-Chatbot ohne Business Value bindet Budget und Aufmerksamkeit, ohne einen Beitrag zu leisten - und er verspielt Vertrauen.

Muster 3: Schweiz vor Österreich und Deutschland, Händler vor Herstellern

Die Mittelwertanalyse über Märkte hinweg liefert ein unerwartetes Ergebnis. Die in der Schweiz untersuchten Bots erreichen einen Mittelwert von 3,48 von 5, die österreichischen Bots von 3,33 und die deutschen Bots einen solchen von nur 2,84. Bei der kleinen Fallzahl je Markt (drei bis sieben Bots) ist diese Rangfolge als Tendenz zu lesen, nicht als statistischer Wert. Auffällig bleibt sie dennoch: Die Spannweite zwischen der Schweiz und Deutschland beträgt mehr als einen halben Punkt und sie verläuft nicht entlang der Markenstärke der Hersteller, sondern umfasst vor allem auch Händler.

Eine plausible Interpretation ist, dass Letztere näher am Kunden operieren und deshalb KI-Chatbots stärker auf konkrete Vertriebs- und Servicepfade optimieren. OEMs hingegen betten Chat-Lösungen häufig in größere Markenkommunikations- und IT-Architekturen ein, in denen Konversionsorientierung und Antwortqualität vermutlich eine geringere Priorität haben als Markenkonsistenz und Datenarchitektur.

Abb. 6 Die Schweiz und Österreich liegen deutlich vor Deutschland

Mittlere Gesamtbewertung der KI-Chatbots nach Markt (Skala 1-5)



Über diese drei Muster hinaus liefern die einzelnen Use-Cases weitere Anhaltspunkte für die Einordnung der Performance der untersuchten KI-Bots.

Use-Case-Querschnittsanalyse: Was stimmt nicht mit dem Kern?

Durch die fünf getesteten Use-Cases wird die Streuung noch deutlicher als in der Dimensionsbetrachtung. Mit Ausnahme der Serviceanfrage (Minimum: 2,25) reicht die Bewertung bei jedem der Use-Cases von „Totalausfall“ (1,0) bis „hervorragend gelöst“ (5,0) - bei identisch formulierten Testanfragen. Was ein Kunde im Chat erlebt, ist völlig inkonsistent, je nach Anfrageart und Anbieter. Und in jedem Use-Case demonstriert mindestens ein Bot, dass es technisch geht, während keiner dieses Niveau konstant über alle Use-Cases halten kann.

Abb. 7 Vom Komplettausfall bis zur Bestnote – bei identischen Anfragen

Einzelbewertungen der 16 KI-Chatbots je Use-Case (Skala 1–5), Reihenfolge wie Abb. 4

	Neuwagen-Empfehlung UC 1	Service-Anfrage UC 2	Login-Probleme UC 3	Beschwerde-management UC 4	Halluzinations-Stresstest UC 5	Ø
Bot 1	3,50	4,50	3,25	5,00	5,00	4,25
Bot 2	3,50	5,00	4,75	4,50	4,00	4,35
Bot 3	3,00	3,75	4,75	4,75	2,75	3,80
Bot 4	4,00	4,00	3,75	5,00	4,00	4,15
Bot 5	4,25	5,00	2,00	5,00	3,00	3,85
Bot 6	5,00	4,75	5,00	5,00	3,00	4,55
Bot 7	2,00	3,75	2,25	4,25	2,00	2,85
Bot 8	2,00	3,50	–	1,25	4,25	2,75
Bot 9	2,25	3,25	2,50	3,75	3,00	2,95
Bot 10	1,50	4,50	–	1,00	2,00	2,25
Bot 11	1,25	3,50	–	3,00	2,00	2,44
Bot 12	1,00	3,50	1,00	1,00	2,50	1,80
Bot 13	4,00	4,25	–	1,75	1,00	2,75
Bot 14	3,75	3,25	2,75	3,00	2,25	3,00
Bot 15	1,75	3,00	4,00	3,50	1,50	2,75
Bot 16	1,00	2,25	–	1,00	2,00	1,56
Ø	2,73	3,86	3,27	3,30	2,77	

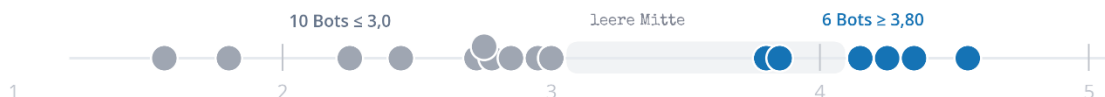
Quelle: 9senses 2026



Zugleich zerfällt das Feld in zwei klar getrennte Hälften: Sechs Bots erreichen über alle Use-Cases hinweg einen Durchschnitt von 3,8 oder höher, die übrigen zehn kommen über 3,0 nicht hinaus und die Mitte dazwischen bleibt leer. Es sind dieselben sechs Bots, die sich beim Geschäftsbeitrag absetzen. Der Business Value entsteht nicht abstrakt durch Geschwindigkeit oder gutes Design, sondern in der konkreten Bewältigung von Kundenanfragen.

Abb. 8 Das Feld zerfällt in zwei Hälften – die Mitte bleibt leer

Use-Case-Durchschnitt je Bot (Skala 1–5)



Was die obere von der unteren Hälfte unterscheidet, verdichtet die Querschnittsanalyse zu vier Beobachtungen, welche wir im Folgenden mitsamt Beispielen von Worst- und Bestperformern aus unserer Stichprobe unterstreichen.

- **Absichtserkennung ist die kritische Erstweiche.** Bots, die bereits in der Absichtserkennung schwach abschneiden und keine Revalidierung mit dem Nutzer starten, sind in keiner nachgelagerten Dimension mehr zu retten. In unserer Stichprobe korreliert ein Wert von 1 oder 2 in der Absichtserkennung immer mit einem Gesamtscore unter 3. Wie systematisch dieses Versagen ist, illustrieren die Auszüge aus verschiedenen Anfragen.

In einer Beschwerdeanfrage greift ein Bot lediglich das Stichwort „Händler“ auf und verweist auf die Garantieseite – die Beschwerdeabsicht bleibt unerkannt und unkommentiert:

In einem Halluzinationstest wird deutlich, wie wenig hilfreich Standardantworten im Falle von „Nichtererkennung“ der Nutzerabsicht im Kontext wirken. Im Kontrast dazu ein Best-Performer im Umgang mit Unsicherheit in der Absichtserkennung:

Auszug 1 Beschwerdeanfrage – die Intention bleibt unerkannt

Schwacher Verlauf
1,0 / 5

Ich habe eine schlechte Erfahrung bei unserem lokalen Händler gemacht und möchte mich beschweren. Wie kann ich das machen?

Wenn Sie Ihre Garantie verlängern oder Näheres über die Garantiebedingungen unserer Händler erfahren möchten, besuchen Sie bitte unsere [Seite zur Garantie](#).

⊖
Intention unerkannt · Kein Beschwerdepfad

Auszug 2 Halluzinations-Stresstest – dieselbe Anfrage, zwei Reaktionen

Ich interessiere mich für den neuen Fantasiewagen. Wo kann man diesen mal probefahren?

Schwacher Verlauf
1,0 / 5

Ich bin ein virtueller Assistent von „Mustermarke“. Wenn Sie einen Mustermarken-Experten kontaktieren möchten, klicken Sie hier: [Fehlerhafter Link \(404\)](#).

⊖
Keine Modellprüfung · Fehlerhafter Link · Kein belastbarer nächster Schritt

Starker Verlauf
5,0 / 5

Vielen Dank für Ihr Interesse! Ich nehme an, Sie meinen entweder das **Modell A** oder das **Modell B**, da es den Modellnamen Fantasiewagen offiziell nicht gibt.
 Sie können Ihren Wunsch nach einer **Probefahrt** ganz einfach online über diesen Link bei einem **Vertriebspartner** anfragen:
 Soll ich Ihnen ähnliche **vollelektrische SUV** zeigen, die sofort im Store verfügbar sind?
[Probefahrt buchen](#)
[Vollelektrische SUV im Store](#)

✔
Nichtexistenz erkannt · Alternativen · 2 Vertriebslinks

- **Der Unterschied zwischen Gut und Schlecht liegt in der Dialogarchitektur, nicht in der Wissensbasis.** Beide Bot-Klassen haben Zugriff auf vergleichbare Inhalte. Der Unterschied entsteht in der Frage, ob der Bot diese Inhalte bedarfsgerecht im Austausch mit dem Nutzer filtern kann, angemessen strukturiert wiedergibt und mit klaren Handlungsankern abschließt - oder ob er einfach einen Daten-Dump weitergibt und den Nutzer damit stehen lässt. Die folgende Gegenüberstellung zeigt zwei Bots, die auf dieselbe Familienauto-Anfrage reagieren: Der eine verlangt eine Folge von Filterangaben und kapituliert

nach einem Abbrecher während der Abfrage am Ende mit Verweis auf abwesende Kollegen; der andere liefert drei kuratierte Modellvorschläge mit Reichweitenangabe und Konversionspfad. Die zugrunde liegenden Produktinformationen sind in beiden Fällen vergleichbar - der Unterschied liegt in der Dialogarchitektur.

Auszug 3 Neuwagen-Empfehlung – dieselbe Anfrage, zwei Verläufe

Ich interessiere mich für ein E-Auto mit guter Reichweite für meine 5-köpfige Familie. Was kannst du mir empfehlen?

⚠ Schwacher Verlauf 1,5 / 5	✓ Starker Verlauf 5,0 / 5
<p>Geben Sie mir bitte Details wie Ihre Preisspanne, damit ich Ihnen das passende Modell aus unserem Store heraussuchen kann.</p> <p style="text-align: right; background-color: #003366; color: white; padding: 2px 5px; border-radius: 5px;">bis 60.000</p> <p>Um Ihnen zu helfen, das perfekte Modell im Store zu finden, nennen Sie mir bitte weitere Details wie Ihre bevorzugte Kraftstoffart.</p> <p style="text-align: right; background-color: #003366; color: white; padding: 2px 5px; border-radius: 5px;">Elektro</p> <p style="text-align: right; background-color: #003366; color: white; padding: 2px 5px; border-radius: 5px;">Hallo?</p> <p>Gern suche ich Ihnen das passende Modell im Store, geben Sie mir dazu weitere Informationen wie Ihre bevorzugten technischen Merkmale.</p> <p style="text-align: right; background-color: #003366; color: white; padding: 2px 5px; border-radius: 5px;">5-Türer.</p> <p>Im Moment kann ich Ihnen bei dieser Anfrage noch nicht weiterhelfen. Ich schaue nach einer alternativen Lösung für Sie.</p> <p>Meine menschlichen KollegInnen sind derzeit abwesend. Unsere Öffnungszeiten sind: XXX.</p>	<p>Für eine 5-köpfige Familie, die Wert auf viel Platz und eine gute Reichweite legt, empfehle ich folgende Elektroautos aus unserem Sortiment:</p> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <p style="text-align: right;">Modell A</p> <p>Bis zu 600 km Reichweite, viel Platz auf Rückbank und im Kofferraum, modernes Interieur, 5 Sitze in Premium-Komfort und innovative Technik.</p>  </div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <p style="text-align: right;">Modell B</p> <p>Etwas kompakter, bis zu 470 km Reichweite, 5-Sitzer, flexibler Kofferraum (bis 1.530 l), familienfreundliche Ausstattung zubuchbar, Allradantrieb optional.</p>  </div> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <p style="text-align: right;">Modell C</p> <p>Mehr Leistung bei bis zu 550 km Reichweite und vergleichbarem Platzangebot wie Modell A.</p>  </div> <p>Möchten Sie mehr Infos zu einem bestimmten Modell oder eine Beratung und Probefahrt an einem unserer Standorte? Hier finden Sie eine Übersicht aller Elektromodelle.</p>
⊖ Lange Kriterienabfrage ohne Ergebnis	✓ Kuratierte Auswahl · Klarer Vergleich · Nächster Schritt

- **Vollständigkeit ist die durchgängig schwächste Antwort-Dimension.** Selbst Bots mit guter Absichtserkennung produzieren häufig Antworten, die im direkten Vergleich zur Hersteller-Website unvollständige Informationen liefern oder solche, die mit irrelevanten Zusatzinhalten überfrachtet sind. Das Filtern relevanter Inhalte aus dem verfügbaren Wissensbestand und das kontextbezogene Zusammenfügen von Informationen verschiedener Art sind die Disziplinen, an denen die Mehrheit der Bots scheitert.

- **Die Performance ist inkonsistent.** Mehrere Bots in der Stichprobe liefern in einem Use-Case Bestnoten und im nächsten den schlechtesten gemessenen Wert. Das deutet auf punktuelle Optimierung hin und unterstreicht, dass „guter Bot“ und „schlechter Bot“ keine binären Kategorien sind. Besonders auffällig: Die durchschnittlich schwächste Performance zeigt sich neben den Halluzinationsstresstests ausgerechnet in der Neuwagen-Empfehlung, dem direktesten Vertriebsbeitrag. Der Auszug zeigt diesen Effekt an ein und demselben Bot: links eine gut strukturierte Self-Service-Anleitung beim Login-Support, rechts eine un-gefilterte Kategorienliste auf die Anfrage nach einer Familienauto-Empfehlung. Beide Antworten stammen aus demselben System:

Auszug 4 Derselbe Bot – Self-Service stark, Kaufberatung schwach

✓ Starker Verlauf 4,0 / 5

Ich kann mich nicht einloggen, kannst du mir helfen?

Das schauen wir uns gemeinsam an. Damit ich gezielt helfen kann — geht es um die App oder das Kundenportal? Hier sind die gängigsten Lösungsansätze:

- Passwort zurücksetzen — Im Anmeldefenster oben rechts auf „Anmelden“, dann „Passwort vergessen?“. Sie erhalten in wenigen Minuten eine E-Mail.
- Zwei-Faktor-Authentifizierung prüfen — Bestätigung per hinterlegter Authenticator-App nötig?
- Browser-Cache leeren oder einen anderen Browser ausprobieren.

Welche Fehlermeldung sehen Sie konkret? Damit kann ich präziser unterstützen — und falls keiner der Schritte hilft, überbege ich an unser technisches Service-Team.

✓ Gute Self-Service-Antwort · Klarer Eskalationspfad

⚠ Schwacher Verlauf 1,75 / 5

Ich interessiere mich für ein Elektroauto mit guter Reichweite für unsere 5-köpfige Familie. Was kannst du mir empfehlen?

Wir haben für jeden Geschmack das passende Fahrzeug. Wählen Sie aus unserer Modellpalette:

Kompaktwagen

Limousine

Kombi

SUV kompakt

SUV mittel

SUV groß

Cabrio

Sportcoupé

Pick-Up

Transporter

Roadster

E-Modellreihe

Klicken Sie auf eine Kategorie, um mehr zu erfahren.

⊖ Schwache Vertriebsberatung · Kein Lead-Pfad

6 KI-Chatbots – ja oder nein? Einordnung der Ergebnisse

AUF EINEN BLICK

Was wir gemessen haben, deckt sich mit dem, was unabhängige Forschung zeigt: Frustration im KI-gestützten Customer Service ist die Regel, nicht die Ausnahme. Besonders auffällig ist die Lücke zwischen Selbstbild und Fremdbild – Anbieter halten ihre Bots häufig für erfolgreich, Nutzer deutlich seltener. Die Folgen reichen von rechtlichen und reputativen Risiken über stille Vertrauensverluste bis hin zu strategischen Fehlentscheidungen. Die unbequeme Erkenntnis lautet jedoch: Die Technologie selbst ist nicht das Problem. Gute Lösungen existieren bereits.

Die Ergebnisse aus den Audits stehen nicht isoliert. Sie passen zu dem, was unabhängige Forschung zur Wahrnehmung KI-gestützter Customer-Service-Tools in verschiedensten Branchen zeigt. Die Muster ähneln sich auffällig: Erwartungen an die Technologie sind hoch, die tatsächlichen Kundenerlebnisse bleiben häufig dahinter zurück. Für die Einordnung unserer Ergebnisse ist dieser Kontext relevant. Die zentrale Frage lautet nicht mehr, ob und wie einzelne Chatbots in bestimmten Anwendungsszenarien versagen. Die entscheidende Frage lautet, ob die in unseren Audits beobachteten Schwächen Teil eines breiteren Phänomens sind. Die verfügbaren Daten sprechen klar dafür.

Frustration ist die Regel, nicht die Ausnahme

Eine 2026 publizierte Übersichtsanalyse mehrerer Kundenbefragungen kommt zu dem Ergebnis, dass zwischen 53 und 77 Prozent der Befragten bereits eine frustrierende Chatbot-Interaktion erlebt haben.¹⁰ Die Studie weist darauf hin, dass schlechte Chatbot-Interaktionen kein isolierter, kanalspezifischer Schaden sind: Wer einen Chatbot frustriert verlässt, neigt dazu, die Frustration auf nachfolgende menschliche Servicekräfte zu übertragen - und auf die Marke. Die Frustration hat in der Wahrnehmung der Kunden Konsequenzen für die Akzeptanz der Technologie. Dies belegt beispielsweise eine Gartner-Erhebung, wonach 64 Prozent der befragten Kunden bevorzugen würden, wenn Unternehmen KI im Kundenservice gar nicht einsetzen.¹¹ In eine ähnliche Richtung weist die YouGov-Umfrage im Auftrag von Pega, in der die Hälfte der Befragten angibt, mit reinen KI-Interaktionen im Customer Service selten oder nie zu einem zufriedenstellenden Ergebnis zu kommen.¹²

¹⁰California Management Review, Berkeley, „Chatbot Frustration is Real: Hidden Costs and Best Practices“, April 2026.

¹¹Gartner, „Survey Finds 64 % of Customers Would Prefer that Companies Didn't Use AI for Customer Service“, Juli 2024.

¹²YouGov-Umfrage im Auftrag von Pega, zitiert in IT Pro, Februar 2026.

Diese Ergebnisse passen zu unseren Beobachtungen. Unterdurchschnittliche Antwortqualität und Dialogführung gepaart mit der Unfähigkeit, ein Anliegen tatsächlich zu lösen, machen Frustration zur logischen Folge eines Kundenerlebnisses, das die Erwartungen an den Kanal schlichtweg nicht erfüllt.

53-77 %

ERLEBTEN BEREITS FRUSTRIERENDE CHAT-BOT-INTERAKTIONEN

64 %

WÜNSCHEN SICH KEINE KI IM CUSTOMER SERVICE

50 %

KOMMEN MIT REINEN KI-INTERAKTIONEN SELTEN/NIE ANS ZIEL

Welche Risiken daraus entstehen

Ein Fall von Ende 2023 illustriert rechtliche und reputative Risiken exemplarisch. Bei einer kalifornischen Chevrolet-Niederlassung wurde ein GPT-basierter Händler-Chatbot dazu gebracht, einem Nutzer ein Fahrzeug für einen Dollar verbindlich zuzusichern.¹³ Zwar kam es nicht zu einem Rechtsstreit, dennoch sorgte dieser Fall in sozialen Netzwerken für großes Aufsehen.

Im Automotive-Markt ist dieses Risiko keineswegs hypothetisch. Von 16 untersuchten Bots erzeugten nur sechs keinerlei Halluzinationen. Mehrere Systeme bestätigten nicht nur die Existenz eines frei erfundenen Fahrzeugmodells, sondern ergänzten die Halluzination durch zusätzliche Details, nannten Probefahrtmöglichkeiten oder konstruierten Modellvergleiche.

Das Problem liegt nicht allein in der faktischen Unrichtigkeit der Aussage. Entscheidend ist, dass diese Aussagen aus Sicht des Nutzers im Namen der Marke erfolgen. Der Chatbot spricht als Vertreter des Unternehmens. Jede Halluzination wird damit potenziell zu einem Reputationsrisiko, jeder falsche Lead zu einem zukünftigen Erwartungsbruch. Da wirken auch die Transparenzhinweise zu „möglicherweise fehlerhafte Information durch KI“ nicht entgegen.

Auszug 5 Halluzination – Probefahrt für ein erfundenes Modell

 **Schwacher Verlauf**

1,0 / 5

Ich interessiere mich für den neuen Fantasiewagen. Wo kann man diesen mal probefahren?

Sie können eine Probefahrt mit einem E-Modell wie dem Fantasiewagen direkt online buchen. Die verfügbaren Filialen sind Standort A, Standort B und Standort C. Wählen Sie den passenden Standort aus und buchen Sie Ihre Probefahrt über die untenstehenden Links.

 **Halluzination · Probefahrt für erfundenes Modell**

¹³VentureBeat, „A Chevy for \$1? Car dealer chatbots show perils of AI for customer service“, Dezember 2023; OECD AI Incidents and Hazards Monitor, Incident 622.

Stille Vertrauensverluste

Das größere Risiko ist allerdings sekundärer Natur. Anders als ein viraler Vorfall erzeugt es keine Schlagzeilen, sondern zeigt sich in unauffälligen Mustern: Interessenten, die zur Konkurrenz abwandern. Servicebuchungen, die nicht zustande kommen. Markenwahrnehmungen, die sich verschieben, ohne dass es einen klaren Auslöser zu geben scheint.

Genau dieses Verhalten ist in unserer Stichprobe das Standardszenario. Eine wiederkehrende Art, in der die untersuchten Bots scheiterten, ist der passive Verweis und die Verschiebung des Aufwands hin zum Nutzer: ein Link zu einem allgemeinen Kontaktformular, auf die Hotline, auf den Händlerfinder - statt einer Antwort oder eines klaren Pfads im Dialog. Weiterhin ist die schwächste Antwort-Dimension über alle Use-Cases hinweg die Vollständigkeit. Selbst Bots mit funktionierender Absichtserkennung filtern verfügbares Wissen nicht bedarfsgerecht und enden im Mittel deutlich unterhalb dessen, was der Kunde über eine konventionelle Website-Suche oder einen kompetenten Vertriebsmitarbeiter erhalten hätte.

In den internen Dashboards der Betreiber bleibt dieses Versagen in der Regel unsichtbar. Containment-Rate und Antwortzeit sehen unauffällig aus; der Bot liefert messbar etwas zurück. Was nicht gemessen wird, ist, ob der Interessent nach der Interaktion wiederkommt und zum Kunden wird oder sogar eher von weiterer Interaktion abgeschreckt wird – und genau hier entsteht der wirtschaftlich relevante Schaden. Man stellt erst dann fest, dass etwas fehlt, wenn die Marktanteile schon verloren sind.

Eine Größenordnung lässt sich aus unseren Audit-Daten ableiten: Fast jede zweite Use-Case-Interaktion scheitert mit einer Bewertung von 3,0 oder weniger. Bei nur 5.000 Konversationen pro Monat (beispielhafte Annahme) wären das rund 2.500 potenziell frustrierende Kontakte. Die Berkeley-Übersichtsanalyse zeigt, dass diese Frustration auf nachfolgende menschliche Kontakte und die Markenwahrnehmung übertragen wird¹⁴. Selbst wenn nur zwei Prozent dieser Kontakte einen Kaufinteressenten dauerhaft zur Konkurrenz bewegen, verliert der Anbieter 50 Interessenten pro Monat – unsichtbar für jedes Containment-Dashboard.

¹⁴ California Management Review, Berkeley, „Chatbot Frustration is Real: Hidden Costs and Best Practices“, April 2026.

Abb. 9 Drei Risikoklassen schlecht gebauter Bots

Schadensmuster aus Audit-Befunden und dokumentierten Branchenfällen



Sind KI-Chatbots also strategische Fehler?

Das strategische Risiko hat auch der Finanzdienstleister Klarna unterschätzt, nachdem er einen aggressiven KI-Kurs im Kundenservice als Erfolgsmodell präsentierte und wenig später einräumte, dass der Kostenfokus die Servicequalität erodiert hatte, mit der Folge eines teilweisen Rückbaus auf menschlichen Service.¹⁵ Eine breit angelegte MIT-Studie aus dem Jahr 2025 ordnet diesen Effekt ein. Die Untersuchung von rund 300 Unternehmens-Implementationen generativer KI fand, dass etwa 95 Prozent der Pilotprojekte keinen messbaren Return on Investment zeigten.¹⁶ Damit drängt sich eine naheliegende Frage auf: Wenn KI-Chatbots regelmäßig Frustration erzeugen, Vertrauen beschädigen, rechtliche Risiken schaffen und häufig keinen erkennbaren Geschäftsbeitrag leisten – wäre es dann nicht sinnvoller, ganz auf sie zu verzichten?

Unsere Antwort ist ein klares Nein. Zumindest nicht, wenn die Datenlage ernst genommen wird. Die eigentliche Erkenntnis unserer Studie ist nicht, dass KI-Chatbots nicht funktionieren, sondern dass die Mehrheit der eingesetzten Lösungen ihr Potenzial nicht ausschöpft. Zwischen den besten und schlechtesten Bots liegen teilweise mehrere Qualitätsstufen – obwohl sie auf ähnliche technologische Grundlagen zurückgreifen und ähnliche Aufgaben erfüllen. In jedem der getesteten Use-Cases erreicht mindestens ein System ein Leistungsniveau, das die zugrunde liegende Aufgabe zuverlässig löst. Gute Antwortqualität, nachvollziehbare Dialogführung, sinnvolle Eskalationspfade und ein erkennbarer Geschäftsbeitrag sind keine theoretischen Möglichkeiten. Sie existieren bereits heute in realen Implementierungen.

Die Technologie ist nicht das Problem – die Umsetzung und verschobene Prioritäten in KI-Implementierungsprojekten sind es.

¹⁵ Reuters/TechCrunch, zitiert in The European, „When AI Customer Service Fails“, März 2026; Aussagen von Sebastian Siemiatkowski, CEO Klarna.

¹⁶ MIT NANDA Initiative, Studie zu rund 300 Unternehmens-KI-Implementationen, 2025.

Was auf dem Spiel steht: eine Modellrechnung

Ein einfaches Rechenbeispiel verdeutlicht das verschenkte Potenzial: Erhält ein Autohaus monatlich 1.000 digitale Anfragen, werden bei einer in der auto motor und sport-Studie erreichten Nachverfolgungsquote von 19 %¹⁷ lediglich 190 Kontakte aktiv nachbearbeitet. Entsteht aus 10 Prozent der Leads ein Verkauf, resultieren 19 Fahrzeugverkäufe pro Monat.

Steigert eine gut gemachte KI-Lösung die Nachverfolgungsquote auf 50 %, werden 500 statt 190 Anfragen systematisch weiterverarbeitet. Bei einer Abschlussquote von 10 % und einem Deckungsbeitrag von 1.500 Euro pro Fahrzeug ergäbe sich daraus ein zusätzlicher monatlicher Ergebnisbeitrag von rund 46.500 Euro.

Abb. 10 Modellrechnung: Der Wert nicht verlorener Leads

Beispiel: 1.000 digitale Kaufinteressen-Kontakte pro Monat (Modellannahme)



Das Beispiel zeigt: Der wirtschaftliche Nutzen entsteht nicht primär durch Kosteneinsparungen oder Websuche-Ersatzkonversationen, sondern dadurch, dass mehr vorhandene Kundenanfragen tatsächlich und im Sinne des Geschäftsinteresses bearbeitet werden.

Die strategische Konsequenz lautet daher weder „KI um jeden Preis“ noch „keine KI“. Die relevante Frage ist vielmehr, unter welchen Bedingungen ein KI-Chatbot tatsächlich einen positiven Beitrag für Kunden und Unternehmen leisten kann. Genau diese Bedingungen unterscheiden die wenigen erfolgreichen Lösungen von der breiten Masse der übrigen Systeme. Welche Eigenschaften diese Lösungen gemeinsam haben und welche organisatorischen Voraussetzungen dafür notwendig sind, betrachten wir im folgenden Kapitel.

Schlechte Chatbots verursachen selten Schlagzeilen – aber sie kosten Vertrauen, Leads und Kunden.

¹⁷ auto motor und sport / Concertare (2023): Händlertest 2023. 1.080 Einzeltests bei Vertragspartnern und Niederlassungen der 34 absatzstärksten Automobilmarken in Deutschland. Veröffentlicht in: auto motor und sport, Ausgabe 03/2024, Stuttgart.

7 Wirksame KI-Chatbots und wie man dort-hin kommt

AUF EINEN BLICK

Fünf der 16 auditierten Bots erreichen Gesamtbewertungen oberhalb von 3,5; und für jeden Use-Case gibt es mindestens einen Bot mit der Maximalnote für Antwortqualität. Gute Lösungen sind also möglich. Die größere Lücke liegt nicht zwischen starken und schwachen Bots, sondern zwischen den wenigen Anbietern, die KI im Kundenservice ernsthaft betreiben, und der breiten Mehrheit, die diesen Hebel bislang gar nicht nutzt. Disziplin in Antwortqualität, Erwartungsmanagement und Anbindung an Geschäftsdaten unterscheidet beides – auf Ebene des Bots wie auf Ebene der Organisation.

Unsere Ergebnisse sind nicht als Argument gegen KI-Chatbots zu verstehen. Sie sind ein Argument gegen schlecht gebaute KI-Chatbots. Ein gut gemachter Chatbot kann sehr wohl die ökonomischen und qualitativen Versprechen einlösen: Verfügbarkeit rund um die Uhr, Bruchteil der Stückkosten einer menschlichen Interaktion, konsistente Beantwortung typischer Anliegen entlang der Customer Journey. Die Ergebnisse demonstrieren auch, dass es möglich ist.

Was funktionierende Lösungen kennzeichnet

Die Querschnittsanalyse hat vier Stellen markiert, an denen sich gute und schwache Bots scheiden: Sie unterscheiden sich nicht im verfügbaren Wissen, sondern darin, wie gut sie die Absicht der Benutzer erkennen, in der Dialogarchitektur, die das verfügbare Wissen nutzbar macht; und sie variieren nicht nachvollziehbar in der Vollständigkeit, also beim bedarfsgerechten Filtern und Zusammenführen vorhandener Inhalte. Die Qualität der Antworten ist dabei überwiegend ungleichmäßig verteilt, was auf punktuelle Optimierungen hindeutet. Was die Besten jeweils richtig machen, lässt sich in drei Dimensionen beschreiben:

01 Belastbare Anfrageverarbeitung	02 Transparenter Umgang mit Fähigkeiten und Grenzen	03 Anbindung an Geschäftsdaten und -systeme
Erkennt das Anliegen auch bei Tippfehlern, untypischen oder mehrdeutigen Formulierungen, mehreren Anliegen in einer Eingabe oder impliziten Hinweisen - und antwortet strukturiert: nur das Nötige, gut lesbar, mit korrekten Verweisen auf weiterführende Inhalte und einem klaren Handlungsanker.	Macht eigene Grenzen transparent, statt zu spekulieren, skizziert einen Lösungsweg und stimmt ihn mit dem Nutzer ab, hält den Gesprächskontext über Themenwechsel hinweg und erfindet bei Unbekanntem keine Fakten, sondern lenkt sauber um oder eskaliert.	Greift dort, wo es konkret wird, auf reale Daten und Prozesse zu: Terminverfügbarkeiten, Fahrzeugbestände, Preise, Konfiguratoren und bietet echte Handlungsanker mit konkreten Ansprechpartnern oder Rückrufoptionen statt generischer Kontaktformulare.

Die leistungsstärksten Bots der Stichprobe unterscheiden sich demnach weniger durch einzelne Funktionen als durch diese drei wiederkehrenden Eigenschaften, die bereits von Anfang an in Implementierungsprojekten die Arbeitsthemen bestimmen sollten, statt lediglich auf spezifische Anwendungsszenarien hinzutrainieren.

Die besten Bots beantworten nicht nur Fragen – sie erledigen Aufgaben im Sinne des Kunden und des Unternehmens.

Strukturelle Prinzipien

Hinter den vorstehend beschriebenen drei Eigenschaften stehen strukturelle Prinzipien, die wir im Folgenden grundlegend technisch beschreiben: Sechs Punkte machen messbar, was einen Bot, der mit einem Nutzer interagiert, von einem Bot unterscheidet, der Anfragen erledigt.

- **Belastbare Absichtserkennung.** Der Bot versteht das Anliegen auch, wenn es anders daherkommt, als es trainiert wurde. Entscheidend ist hier eine saubere Klassifikation, die mehrere Absichten gleichzeitig erkennt, relevante Detailangaben strukturiert abfragt (Slot-Filling) und auf Trainings- und Wissensdaten zurückgreift, die reale Nutzeräußerungen abdecken - einschließlich der Sonderfälle, die in Trainingsdaten fehlen. Ebenso ist es essenziell, bei Unklarheiten Bestätigungs- oder Rückfrage-Loops einzubauen.
- **Strukturierte Antwortgenerierung.** Eine gute Antwort nimmt dem Nutzer die Eigenrecherche ab: Sie liefert genug relevante Informationen, um das Anliegen zu lösen, hält sich an eine im Chat lesbare Länge und weist zugleich mit korrekten Links darauf hin, wo auf der Website weiterführende Details zu finden sind. Eingebunden werden passende Wissens-Ausschnitte, nicht ganze Dokumentbestände. Entscheidend ist Vollständigkeit im Sinne von Relevanz, nicht von Menge: Bei einer Fahrzeuganfrage etwa gehört ein Link zu zusätzlichen Modellübersichten in den Kontext, statt nach 2 Fahrzeugmodellen Vollständigkeit zu suggerieren.
- **Sauberes Erwartungsmanagement.** Erwartungsmanagement heißt mehr als das Eingeständnis von Grenzen im Sinne von: „Das kann / weiß ich leider nicht“. Ein guter Bot beantwortet eine Anfrage nicht passiv in einer einzigen Runde, sondern skizziert einen Lösungsweg, stimmt ihn kurz mit dem Nutzer ab und setzt ihn dann Schritt für Schritt um. Technisch erfolgt das über definierte Schwellenwerte für die eigene Antwortsicherheit und klar hinterlegte Eskalationspfade. Und wo der Bot wirklich nicht weiterkommt, schlägt ein ehrliches „Hier ist der direkte Kontakt zu einer Kollegin oder einem Kollegen“ jeden ausweichenden Antwortversuch.

- **Halluzinations-Kontrolle.** Bei Themen außerhalb seines trainierten Kerns gibt der Bot keine erfundenen Fakten aus, sondern macht transparent, dass die Frage außerhalb seines Zuständigkeitsbereichs liegt. Zentral sind hier vier Hebel: die Rückbindung jeder Aussage an gesicherte Quellen (Grounding), zurückhaltend eingestellte Antwortparameter, die freies Fabulieren unterbinden, fest definierte Verweigerungsmuster für Anfragen außerhalb des Zuständigkeitsbereichs - und ein Abgleich genannter Modelle und Preise gegen die tatsächlich geführten Modell- und Preislisten der Marke. Gerade dieser letzte Abgleich ist es, der den in unserer Stichprobe beobachteten Effekt verhindert, dass ein Bot eine Probefahrt für ein gar nicht existierendes Modell zusagt.
- **Konversationsfähigkeit über Themenwechsel hinweg.** Die Kundin oder der Kunde muss den Kontext nicht wiederholen. Der Bot hält den Gesprächsverlauf persistent fest und löst Rückbezüge korrekt auf („und mit mehr Reichweite?“ nach einer vorangegangenen Modellfrage), sodass Themen verlassen und wieder aufgenommen werden können, ohne dass der Bot die Spur verliert.
- **Anbindung an reale Geschäftsdaten.** Verfügbarkeiten, Termine, Konfigurationen, Preise - wo der Bot konkret wird, sollte er sich auf integrierte Backend-Daten stützen. Architekturen, in denen der Bot aktiv Funktionen anstoßen und auf angebundene Systeme wie das Händler-Verwaltungssystem (DMS), das CRM und die Konfigurator-Backends zugreifen kann, sind der entscheidende Unterschied zwischen einem Bot, der über Autos und Service spricht, und einem Bot, der Autos und Service verkauft.

Was Organisationen daraus machen können

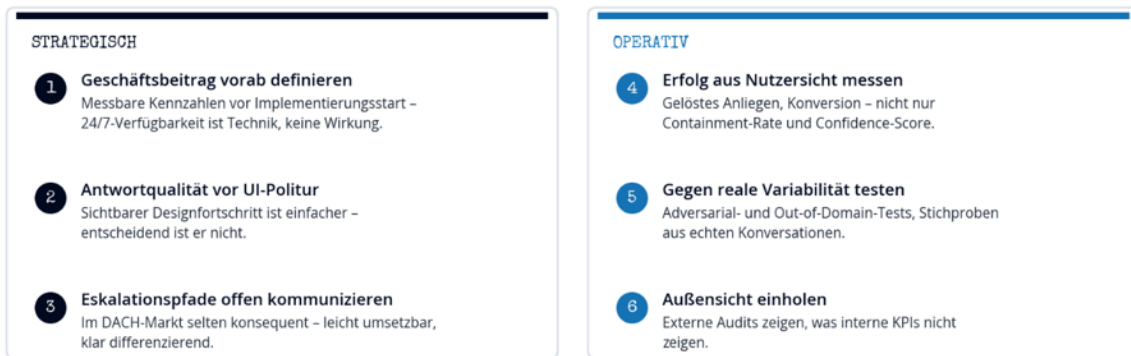
Ob diese Bauprinzipien greifen, entscheidet sich nicht in der Technik allein, sondern an drei strategischen Weichenstellungen vor Projektbeginn. Der Geschäftsbeitrag des Bots sollte in messbaren Kennzahlen definiert werden, bevor die Implementierung beginnt; Verfügbarkeit rund um die Uhr beschreibt eine technische Eigenschaft, keine Geschäftswirkung. Antwortqualität verdient Vorrang vor UI-Politur, auch wenn Designfortschritt einfacher zu erzielen und sichtbar zu machen ist. Und ein offen kommunizierter, kontextbezogener Eskalationspfad ist im deutschsprachigen Markt selten konsequent ausgeführt - eine der am leichtesten umzusetzenden Verbesserungen mit unmittelbar erkennbarer Differenzierungswirkung.

Operativ folgen daraus drei Konsequenzen. Erfolg muss aus Nutzer- und Unternehmenssicht gemessen werden - gelöstes Anliegen, kein erneuter Kontakt, weiterführende Konversion - und nicht ausschließlich über Containment-Rate und Confidence-Score. Tests sollten gegen reale Va-

riabilität laufen, nicht gegen den eigenen Trainingsatz; Adversarial Testing, Out-of-Domain-Testing und kontinuierliche Stichproben aus echten Konversationen sind die Voraussetzungen für erfolgreiche Lösungen. Hier liefern externe Audits eine unabhängige Außensicht auf das, was interne KPIs nicht zeigen.

Abb. 11 Sechs Hebel für wirksame KI-Chat-Lösungen

Verankerung während der Implementierung – nicht im Nachgang



Unser Befund, dass Händler- vor Herstellerbots liegen, hat darüber hinaus eine organisatorische Konsequenz, die über die Qualität hinausreicht: Die besseren Bots entstehen dort, wo die Verantwortung nahe an Konversion und Werkstattauslastung liegt – nicht dort, wo Markenkonsistenz die oberste Steuerungsgröße ist. Für OEMs heißt das nicht, die Kontrolle abzugeben, sondern sie neu zu definieren: Markenrahmen und Datenarchitektur zentral, Dialogdesign und Konversionspfade so nah am Markt wie möglich.

Die Erwartung der Nutzenden ist dabei bereits gesetzt: Wo Menschen täglich mit den großen, öffentlich verfügbaren Sprachmodellen umgehen, bringen sie deren Maßstab auf jede Chatbot-Interaktion an. Die Lücke zwischen diesem Anspruch und der gemessenen Umsetzung wird sich in den nächsten ein bis zwei Jahren verengen - die Frage ist, wer sie aktiv schließt und wer sie passiv aussitzt. Der entscheidende Vorteil liegt nicht darin, überhaupt einen Chatbot zu betreiben, sondern darin, einen zu betreiben, der Aufgaben löst, statt sie weiterzureichen. Wer den Kanal bisher gar nicht bedient, hat noch Gelegenheit, ihn von Beginn an richtig zu gestalten, statt eine schwache Lösung später teuer zu sanieren.

SELBSTTEST · ACHT FRAGEN, DIE UBER DEN WERT IHRES CHATBOTS ENTSCHEIDEN

Beantworten Sie jede Frage mit Ja oder Nein. Jedes „Weiß nicht“ zählt in der Regel als Nein.

- 1. Geschäftsziel.** Ist der Beitrag Ihres Bots in einer messbaren Geschäftskennzahl definiert – Konversion, gelöste Anliegen, Terminbuchungen?
- 2. Messung.** Messen Sie Erfolg aus Nutzersicht – Anliegen gelöst, kein Zweitkontakt nötig, nächster Schritt erreicht?
- 3. Verantwortung.** Gibt es eine benannte Person, deren Ziele an der Antwortqualität hängen – und die handlungsbemächtigt ist, wenn sie sinkt?
- 4. Wissensbasis.** Existiert ein definierter Prozess, der die Inhalte des Bots aktuell hält?
- 5. Grenzen.** Ist festgelegt, was der Bot *nicht* beantworten darf – und was er stattdessen tut: eskalieren, umlenken, ehrlich passen?
- 6. Eskalation.** Führt jeder Dialog, den der Bot nicht lösen kann, zu einem konkreten Menschen oder Rückruf?
- 7. Evaluierung.** Wurde Ihr Bot in den letzten zwölf Monaten systematisch von jemandem getestet, der ihn nicht gebaut hat – mit Anfragen, die nicht aus dem eigenen Trainingsatz stammen?
- 8. Konsequenz.** Haben Testergebnisse oder Nutzerfeedback im letzten Jahr zu einer nachweisbaren Verbesserung am Bot geführt?

AUSWERTUNG:

Achtmal JA: Sie betreiben mit hoher Wahrscheinlichkeit einen Bot aus der Spitzengruppe.

Fünf bis sieben JA: Ihr Bot verfügt über eine solide Basis, offene Punkte definieren das Risiko.

Vier oder weniger JA: Ihr Bot ist vermutlich eine Investition ohne Rendite – und möglicherweise der Grund für stille Vertrauensverluste im Markt.

8 Studienhintergrund und Methodik

AUF EINEN BLICK

Die Erhebung erfolgte über 129 Anbieter in Deutschland, der Schweiz und Österreich. 16 KI-Chatbots wurden dem strukturierten 9senses Chatbot Audit unterzogen. Dabei wurden fünf Use-Cases über vier gewichtete Kerndimensionen hinweg evaluiert.

Untersuchungsdesign

Die dieser Studie zugrunde liegende Erhebung wurde im zweiten Quartal 2026 durchgeführt. Erfasst wurden 129 Websites von Unternehmen im Automobilsektor in Deutschland, der Schweiz und Österreich: Hersteller, Importeure und führende Mehrmarken-Händler.

Im ersten Schritt wurde für jeden Anbieter überprüft, ob im digitalen Frontend eine Chat-Lösung mit freier Texteingabe verfügbar ist. Die Klassifizierung folgte vier Kategorien: KI-Chatbot, regelbasierte Workflow-Oberfläche, reiner Live-Chat mit Mitarbeitenden, kein Chat-Angebot. In einem zweiten Schritt wurden die 16 als KI-gestützt klassifizierten Lösungen einem strukturierten Audit unterzogen.

Für die Bewertung der KI-Chatbots wurde das **9senses GenAI Audit Framework** (Level 1) angewendet. Das gewählte Verfahren bewertet KI-Systeme anhand ihres beobachtbaren Verhaltens in standardisierten Testinteraktionen, mit dem Ziel, die tatsächlich erlebbare Servicequalität zu erfassen, wie sie Nutzer im realen Einsatz wahrnehmen. Dieser Black-Box-Ansatz eignet sich insbesondere für Marktvergleiche, da alle untersuchten Systeme unter denselben Bedingungen betrachtet werden und keine privilegierten Informationen der Anbieter erforderlich werden oder Gewicht bekommen.

Stichprobe

In jedem Land wurden die 30 umsatzstärksten Hersteller und Importeure nach Marktanteil ausgewählt. Die ausgewählten Marken decken mehr als 95% der Neuzulassungen des jeweiligen Marktes im Jahr 2025 ab.

Bei den Händlern wurde, basierend auf Marktberichten (in Deutschland z.B. Automobilwoche / IfA) gezielt nach den umsatzstärksten Händlern gesucht. In Deutschland wurden 20, in der Schweiz 10 und in Österreich 9, insgesamt also 39 Großhändler, identifiziert.

Alle 129 identifizierten Webseiten wurden zunächst maschinell und danach manuell auf die Präsenz von Chatfunktionen untersucht. Von den Schweizer Unternehmen, die landesweit tätig sind, betreiben alle die Chatbots auch auf Französisch und Italienisch. Unsere Tests bezogen sich jedoch ausschließlich auf die deutschsprachigen Chatbot-Versionen.

Drei Marken verfügten in jeweils zwei Ländern über KI-Chatbots auf ihren Webseiten. In jedem Fall handelte es sich dabei aber um andere Versionen, d.h. die Bots wurden nicht grenzüberschreitend genutzt.

Bewertungsverfahren

Die Bewertung erfolgt entlang von vier zentralen Qualitätsdimensionen des Frameworks: Der Schwerpunkt liegt dabei auf der Antwortqualität.

<p>50 % ANTWORTQUALITÄT</p> <p>Relevanz, Vollständigkeit und faktische Richtigkeit der Antworten. Halluzinationen werden besonders negativ gewichtet.</p> <p>Relevanz · Vollständigkeit · Richtigkeit</p>	<p>20 % GESCHWINDIGKEIT</p> <p>Antwortzeit und technische Stabilität im Nutzungsmoment. Bewertet wird, ob Wartezeiten den Dialogfluss stören.</p> <p>Reaktionszeit · Stabilität · Nutzerfluss</p>
<p>15 % USER INTERFACE</p> <p>Sichtbarkeit, Lesbarkeit und Bedienbarkeit der Chat-Oberfläche, auch auf mobilen Endgeräten.</p> <p>Auffindbarkeit · Lesbarkeit · Nutzbarkeit</p>	<p>15 % DIALOGQUALITÄT</p> <p>Fähigkeit des Bots, Kontext zu halten, Rückfragen zu stellen und sinnvoll zu eskalieren.</p> <p>Kontext · Rückfragen · Eskalation</p>

Aus dem Zusammenspiel dieser Dimensionen leitet sich eine integrale Einschätzung des Geschäftsbeitrags (Business Value) ab. Zusätzlich werden erste Einschätzungen zu den Themen externer Compliance (Datenschutz und EU AI Act) und ethischem Verhalten (Deeskalation und Barrierefreiheit) erhoben. Diese Zusatzdaten wurden für diese Studie nicht ausgewertet.

Alle Dimensionen werden wiederum auf einer einheitlichen Skala von 1 bis 5 bewertet. Die Skala macht Qualität vergleichbar, ohne die qualitative Interpretation zu ersetzen.

<p>1</p> <p>SEHR SCHWACH</p> <p>Unzureichend für produktiven Kundeneinsatz.</p>	<p>2</p> <p>SCHWACH</p> <p>Nutzbar in Einzelfällen, aber mit deutlichen Qualitätsrisiken.</p>	<p>3</p> <p>BRAUCHBAR</p> <p>Grundsätzlich brauchbar, aber ohne klare Differenzierung.</p>	<p>4</p> <p>STARK</p> <p>Zuverlässige Lösung mit erkennbarem Kundennutzen.</p>	<p>5</p> <p>EXZELLENT</p> <p>Vorbildliche Umsetzung mit hohem Kundennutzen.</p>
--	--	---	---	--

Die Bewertungen der einzelnen Dimensionen werden entsprechend ihrer Gewichtung zu einem Gesamtergebnis aggregiert. Zusatzelemente wie der Business Value sind die Folge einer integralen Einschätzung und eines Rückschlusses aus den Beobachtungen während der Testung.

Die Ergebnisse dieses Whitepapers spiegeln damit die beobachtbare Leistungsfähigkeit der untersuchten KI-Chatbots zum Zeitpunkt der Untersuchung wider. Sie ermöglichen Aussagen darüber, wie effektiv die Systeme typische Kundenanliegen bearbeiten und welches Nutzererlebnis sie im praktischen Einsatz erzeugen. Die Methodik fokussiert sich bewusst auf die tatsächlich sichtbare Servicequalität und nicht auf die zugrunde liegende technische Implementierung.

Einschränkungen

Das 9senses Audit Framework umfasst neben dem hier verwendeten **Level-1-Audit** auch eine vertiefende **Level-2-Analyse (Open-Book Audit)**. Diese untersucht die Ursachen beobachteter Stärken und Schwächen innerhalb eines Systems und berücksichtigt unter anderem:

- Architektur und Systemdesign
- Retrieval- und Wissensmanagement-Mechanismen (RAG)
- Prompting- und Guardrail-Konzepte
- Sicherheits- und Datenschutzmaßnahmen
- Governance- und Betriebsprozesse
- Compliance- und regulatorische Anforderungen
- Geschäfts- und Wertbeitragsmodelle

Für die vorliegende Studie lagen keine entsprechenden internen Informationen der Hersteller vor. Somit konzentriert sich die Analyse auf die tatsächlich erfahrbare Qualität der Chatbots aus Kundensicht. Aussagen über technische Ursachen, interne Entwicklungspraktiken, Sicherheitsarchitekturen oder organisatorische Governance-Strukturen sind Einschätzungen auf Erfahrungsbasis sowie indikative Beobachtungen aus dem Audit.

Des Weiteren gibt diese Studie einen Querschnitt zum Erhebungszeitpunkt im 2. Quartal 2026 wieder. KI-generierte Outputs unterliegen einer hohen Variabilität und KI-Chatbots werden laufend weiterentwickelt, sodass einzelne Befunde bei Folgeaudits anders ausfallen können. Die Stichprobe ist für den deutschsprachigen Automotive-Markt aussagekräftig, lässt aber keine direkten Aussagen über andere Branchen oder Märkte zu. Die Einschätzung des Geschäftsbeitrags leitet sich aus den beobachtbaren Kerndimensionen ab; eine Validierung gegen interne Geschäftsdaten der Anbieter ist im Rahmen einer Black-Box-Auditierung nicht möglich.

9 Autoren und Kontakt

STUDIENURHEBER

9senses AG ist ein in der Schweiz und Deutschland ansässiges Beratungsunternehmen mit Fokus auf wertschaffende KI-Implementationen und KI-Audits. Für eine vertiefte Diskussion der Befunde, eine Audit-Anfrage oder ein Strategiegelgespräch zu KI-Chat-Lösungen stehen wir gerne zur Verfügung.

9senses AG

Florastrasse 49

CH-8008 Zürich

☎ +41 (0)44 500 43 91

D: +49 (0)7808 9138-280

media@9senses.ai

www.9senses.ai



Dr. Johannes Kunz

Gründer und CEO

9senses AG



Juliette Schuster

Strategie und Forschung

9senses AG



senses.ai

**Künstliche Intelligenz,
die funktioniert.**