Generative KI im kanalübergreifenden Kundendialog

Know-how Der kanalübergreifende Kundendialog mit KI erfordert vor allem eine durchgängige Architektur, die Kontext, Daten und menschliche Kontrolle verbindet.

Von Josef Novak

Der kanalübergreifende Kundendialog wurde durch generative KI und Large Language Models (LLMs) nicht grundlegend verändert, zumindest nicht unabhängig von den Entwicklungen im maschinellen Dialog einzelner Kanäle. Das Kernziel besteht nach wie vor darin, einen kohärenten, logischen Austausch mit dem Kunden über verschiedene Kommunikationskanäle hinweg aufrechtzuerhalten, ohne den roten Faden zu verlieren. Moderne LLMs vereinfachen jedoch die Steuerung und Anpassung von Chatbots innerhalb und über bestimmte Kanäle hinweg erheblich.

Ein zentraler Aspekt dieser Evolution ist die fortschreitende Multimodalität: Multimodale LLMs können während einer laufenden Interaktion in einem klassischen Textkanal problemlos komplexe multimodale Inhalte wie Links, strukturierte Erklärungen und Visualisierungen bereitstellen. Ohne dass hierfür tiefgreifende architektonische Änderungen vorgenommen werden müssen, kann dasselbe zugrundeliegende System diese Antworten so umformulieren, dass sie für einen rein sprachbasierten Kanal wie ein Telefonat geeignet sind. Bei dieser Entwicklung geht es somit weniger um eine vollständige Neukonzeption des kanalübergreifenden Dialogs an sich, sondern primär um die Etablierung effektiverer, kanalbewusster Dialogstrukturen. Das strategische Ziel moderner Architekturen ist es daher, Systeme zu entwickeln, die sich während eines Gesprächs spontan an die Stärken und Grenzen der jeweils aktiven Kommunikationsform anpassen können. Um dies voll auszuschöpfen, ist es wichtig, dass Unternehmensarchitekturen einen einheitlichen Plattform-Ansatz verfolgen. Dies trägt dazu bei, dass die Prozesse vor, während und nach dem Gespräch gemeinsam optimiert werden können – für bestmögliche Customer Journeys und hochgradig effiziente, integrierte Workflow-Updates.

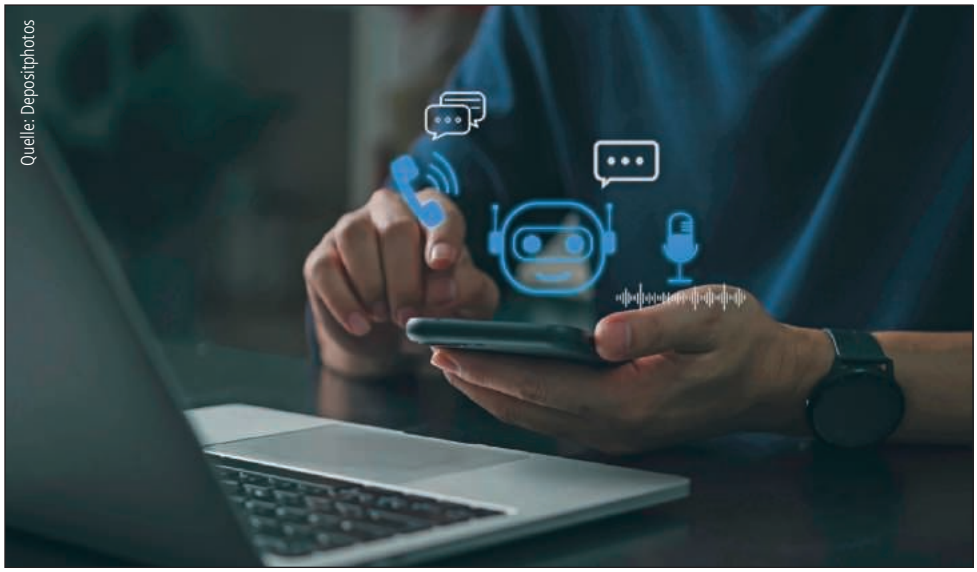
Technische Hürden und die Architektur der nahtlosen Kontinuität

Beim Wechsel eines Kunden zwischen verschiedenen Interaktionsmedien – etwa von einem telefonischen Sprachbot zu einem Chatbot in einer mobilen Applikation – stehen IT-Infrastrukturteams und CX-Verantwortliche vor erheblichen technischen Herausforderungen. Historisch gewachsene Strukturen führen in der Praxis häufig dazu, dass diese Kommunikationsbereiche isoliert voneinander verwaltet werden. Die Gründe hierfür liegen oft in der Nutzung unterschiedlicher Drittanbieter für einzelne Modalitäten, in differierenden Datenformaten der Inhalte oder darin, dass Kunden von vornherein nur einen spezifischen Kanal angefordert haben. Während sich die grundlegende Harmonisierung über ein einheitliches Inhaltsformat lösen lässt und hierfür noch keine spezialisierte KI zwingend erforderlich ist, gestaltet sich die kanalübergreifende Verknüpfung historischer Gesprächsverläufe weitaus komplexer.

Für eine erfolgreiche Zusammenführung dieser Kommunikationsströme müssen in der Enterprise-Architektur mehrere Kernkomponenten implementiert werden:

- **Einheitliche Identitätskopplung:** Es bedarf einer konsistenten Methode, um die Interaktionen innerhalb der Applikation systemübergreifend zu verknüpfen – typischerweise über eine einheitliche Kunden-ID, welche Telefonnummern, E-Mail-Adressen, App-Identitäten und andere relevante Kontoinformationen miteinander verbindet.
- **Kontextuelle Relevanzprüfung:** Das System muss präzise bestimmen können, ob und wie die aktuelle Interaktion, unabhängig vom Medium oder der Modalität, mit früheren Interaktionen zusammenhängt. An dieser Stelle rückt der Einsatz von KI ins Zentrum der Systemarchitektur.
- **Dynamische Historienmodelle:** Idealerweise greift die Umgebung auf ein sich kontinuierlich weiterentwickelndes Modell der relevanten Interaktionshistorie des Kunden zurück, wobei stets ein angemessener Rahmen für Datenschutz und Einwilligung gewahrt bleiben muss.
- **Thread-Differenzierung:** Es muss exakt ermittelt werden, ob eine bestimmte Konversation eine direkte Fortsetzung eines bestehenden multimodalen Threads darstellt oder lediglich Teil der umfassenderen Kundenhistorie ist.

Um die Konversation angemessen und ohne Informationsverlust fortzuführen, ist ein präzises Management des Kontextfensters im LLM unerlässlich. Dieses Kontextfenster muss strukturell so aufgebaut sein, dass es verschiedene Informationsebenen gleichzeitig verarbeitet: einen Hintergrundtext zur vollständigen Kun-


Quelle: Depositphotos

Moderne KI kann den Kundendienst über Chat, Telefon und App unterstützen. Dafür braucht es aber klare Regeln, geschützte Daten und menschliche Kontrolle.

denidentität und -historie, eine verdichtete Zusammenfassung des aktuellen Gesprächsverlaufs sowie die letzten präzisen Gesprächsrunden über alle bisher genutzten Modalitäten hinweg.

Zudem müssen IT-Verantwortliche die spezifische Dynamik des Modalitätswechsels algorithmisch berücksichtigen. Findet ein Wechsel vom Chat zur Sprache statt, muss das KI-System in der Lage sein, auf alle zuvor im Chat geteilten visuellen oder strukturierten Elemente – wie Bilder, Videos, Links oder tabellarische Inhalte – sprachlich präzise einzugehen. Erfolgt der Wechsel umgekehrt von der Sprache zum Chat, verschieben sich die Anforderungen ebenfalls: Da die KI nun über eine schriftliche Aufzeichnung verfügt, der Kunde sich jedoch primär auf sein Gedächtnis stützen muss, gilt es im Chat gezielt auf weitere Details zu verweisen oder diese explizit zu bestätigen.

Menschliche Autonomie im KI-Ökosystem

Im Rahmen eines KI-gesteuerten Ökosystems im Kundenservice stellt sich für das Management die Frage nach der genauen Positionierung des menschlichen Mitarbeiters und den Auswirkungen auf die gesamte Employee Experience (EX). Die strategische Ausrichtung sollte hierbei klar definiert sein: Menschliches Handeln und nicht ausschliesslich die Funktion des menschlichen Agenten besitzt weiterhin höchste Priorität. Das Ziel moderner KI-Architekturen im Enterprise-Umfeld besteht darin, menschliche Anliegen zu fördern sowie menschliche Akteure gezielt zu stärken und zu unterstützen, anstatt Arbeit und Verantwortung unreflektiert auf ungetestete Automatisierungssysteme zu verlagern.

Daraus ergibt sich für die Organisation von Contact Centern, dass menschliche Agenten oder Manager stets die volle Kontrolle sowie die Verantwortung behalten müssen. Alle kritischen Aktionen, die durch eine KI initiiert werden, sollten unter der kontinuierlichen Überwachung, Verwaltung und expliziten Genehmigung durch den Menschen stehen. Die Technologie fungiert somit als Katalysator zur Steigerung der menschlichen Autonomie und Produktivität. Dies wird realisiert, indem den Mitarbeitern präzisere Informationen, optimierte Zusammenfassungen, fundierte Empfehlungen und leistungsfähigere Werkzeuge zur Bewältigung komplexer, kanalübergreifender Interaktionen bereitgestellt werden.

Obwohl KI-Agenten im modernen Kundenservice fest etabliert und nicht mehr wegzudenken sind, ist eine autonome Verantwortung für geschäftskritische Entscheidungen durch diese Systeme auf absehbare Zeit nicht zu empfehlen. Die Erfahrungswerte der Praxis zeigen hier ein klares Bild: Werden unzureichend gesteuerte KI-Systeme mit folgenschweren Entscheidungen, verbindlichen Verpflichtungen oder direkten Kundenergebnissen betraut, treten schnell sichtbare Kettenreaktionen von Fehlern auf. Der strategische Ansatz der kollaborativen KI zielt im Kern darauf ab, ebensolche kaskadierenden Ausfallszenarien systematisch zu verhindern, indem die primäre Rolle der künstlichen Intelligenz konsequent darauf ausgerichtet wird, das menschliche Handeln zu unterstützen und zu fördern. Eine solche komplementäre Architektur führt schliesslich zu sichereren, zuverlässigeren und effektiveren Ergebnissen – gleichermassen für Kunden, Unternehmen und Mitarbeiter.

Dialektvielfalt im Schweizer Markt

Für Unternehmen, die im Schweizer Markt agieren, ergeben sich beim Einsatz von Dialogsystemen spezifische Anforderungen an die Mehrsprachigkeit und den Umgang mit Dialekten innerhalb einer kanalübergreifenden Umgebung. Ein nachhaltiges System muss in der Lage sein, die Erwartungen der lokalen Gemeinschaft präzise zu erfüllen. In der Schweiz bedeutet dies konkret, dass dialogorientierte KI-Systeme eine fachkundige Unterstützung für das exakte Verständnis der verschiedenen lokal gesprochenen Schweizerdeutschen Dialekte bieten müssen. Darüber hinaus sind – sofern vom Kunden oder der Marke gewünscht – entsprechende Text-to-Speech(TTS)- sowie Sprach-zu-Sprache-Antwortfunktionen direkt im jeweiligen Dialekt bereitzustellen.

In den schriftlichen Kanälen wie Chat oder Textnachrichten stellt sich die Situation differenzierter dar: Obwohl moderne KI-Systeme Schweizerdeutsche Dialekte auch in geschriebener Form problemlos verstehen können, existiert für das Schweizerdeutsche keine standardisierte Schriftsprache. Aus diesem technologischen Grund konzentrieren sich textbasierte Antworten in Chatbot-Interaktionen in der gesamten Region typischerweise auf das Standarddeutsche respektive Hochdeutsch.

Für die Systemarchitektur ist es von zentraler Bedeutung, dass die KI den Kunden in seiner bevorzugten Kommunikati-

onsstrategie flexibel versteht, die Antwort jedoch in jener Form generiert, die für den jeweiligen Kanal, die Identität der Marke und die Erwartungshaltung des Kunden am besten geeignet ist. Im Rahmen von Sprachinteraktionen erfordert dies eine dedizierte Dialektunterstützung, während es in schriftlichen Kanälen bedeutet, eine dynamische Mischung aus unterschiedlichen Eingabestrategien der Nutzer präzise zu erkennen und gleichzeitig konsistent in Hochdeutsch zu antworten.

Datensouveränität und Governance unter Schweizer Richtlinien

Die Einhaltung der strengen Schweizer Datenschutzbestimmungen wirft für CIOs und IT-Entscheider die Frage auf, wie globale KI-Innovationen effizient genutzt werden können, ohne die lokale Datenhoheit zu gefährden. Die Marktrealität zeigt jedoch, dass viele der grössten und leistungsfähigsten Sprachmodelle bereits lokal angeboten werden – in zahlreichen Fällen direkt innerhalb der Schweiz und in fast allen verbleibenden Szenarien über spezialisierte Rechenzentren innerhalb der Europäischen Union. Für Unternehmen entfällt dadurch der vermeintliche Widerspruch zwischen dem Zugriff auf globale KI-Innovationen und der strikten Einhaltung lokaler Datenschutz- und Residenzanforderungen.

Der technologische Schlüssel liegt darin, Datensouveränität von Beginn an als eine fundamentale Frage der Systemarchitektur und der Corporate Governance zu behandeln. Die Verarbeitung sensibler Kundendaten, die Speicherung, die Systemprotokollierung (Logging), die Datenaufbewahrung, die Zugriffskontrollen sowie die Modellinteraktionen müssen von Grund auf unter Berücksichtigung der relevanten schweizerischen und europäischen regulatorischen Anforderungen konzipiert werden.

In vielen Infrastrukturbereichen bleiben die reine Speicherung und die standardmässigen Datenprozesse weitgehend identisch mit den Strukturen, die vor den Fortschritten im Bereich der Large Language Models etabliert waren. Die wesentliche Veränderung liegt jedoch darin, dass Unternehmen heute wesentlich präziser und granularer definieren müssen, wo genau die Modell-Inferenz stattfindet, welche spezifischen Daten zur Verarbeitung an das Modell übermittelt werden, wie lange sie dort aufbewahrt werden dürfen und ob diese Daten für Modell-Anpassungen oder ein erneutes Training der Basismodelle herangezogen werden können. Sofern IT-Verantwortliche bei der Bereitstellung und Konfiguration der Systeme die notwendige Sorgfalt walten lassen, können Unternehmen die neuesten technologischen Innovationen vollumfänglich ausschöpfen und gleichzeitig die lokalen Vorgaben an die Datenresidenz und Data-Governance lückenlos einhalten.

Neue Performance-Indikatoren

Die Bewertung des Erfolgs von KI-Systemen im Kundendialog erfordert ein Umdenken hinsichtlich der traditionellen Kennzahlen, da sich in der Praxis der wachsende Bedarf an erweiterten Messgrössen deutlich abzeichnet. Die klassische Average Handling Time (AHT) behält zwar weiterhin ihre Relevanz, wird jedoch nicht mehr als einfache, rein zeitbasierte Dauer-Kennzahl herangezogen, sondern erfordert eine komplexe, mehrdimensionale Berechnung.

Während das übergeordnete Ziel nach wie vor darin besteht, operative Kosten zu minimieren und gleichzeitig die Kundenzufriedenheit zu maximieren, setzen sich die Kostenkomponenten in modernen LLM-Architekturen neu zusammen: Sie umfassen nunmehr die gezielte Modellauswahl, die Token-Nutzung, das Szenariodesign sowie die spezifischen Eskalationsstrategien. Die präzise Auswahl des passenden Szenarios, des am besten geeigneten Modells und des dazugehörigen Token-Budgets erweist sich hierbei als betriebswirtschaftlich entscheidend. Während der Einsatz eines sehr grossen Modells für komplexe und hochwertige Kundeninteraktionen vollkommen gerechtfertigt sein kann, ist er für einfache FAQ-Abfragen oder reine Routing-Aufgaben ökonomisch unnötig.

Parallel dazu wächst in Unternehmen das strategische Interesse, das Contact Center von einer reinen Kostenstelle in eine aktive Ertragsquelle zu transformieren. Dieser Wandel kann durch gezielte Vertriebskennzahlen vorangetrieben werden, die direkt auf KI-basierten Erkenntnissen und Analysen beruhen. Praktische Instrumente hierfür sind automatisierte Hinweise an menschliche Agenten, präzise Empfehlungen für die nächsten optimalen Handlungsschritte (Next-Best-Action) oder zentrumsübergreifende Blitzkampagnen. Diese Mechanismen tragen dazu bei, die Umsatzgenerierung im Contact Center strategisch zu forcieren, und stellen ein wesentliches Feld für die wertschöpfende Zusammenarbeit zwischen menschlichen Mitarbeitern und KI-Agenten dar.

Da menschliche Agenten im Zuge dieser technologischen Unterstützung zunehmend an operativer Autonomie gewinnen, müssen sich auch die Evaluierungsmetriken entsprechend anpassen. Relevante neue Messgrössen umfassen:

- Die exakten Abschlussraten von spezifischen Tools oder Aufgaben durch die Agenten selbst.
- Die qualitative Bewertung von Eskalationsprozessen (Eskalationsquality) an den Schnittstellen zwischen Bot und Mensch.
- Die messbaren Trends zur kontinuierlichen Verbesserung der gesamten Customer Journey über einen längeren Zeitraum.

Hierzu gehört beispielsweise das systematische Tracking der Frage, wie häufig und mit welcher Präzision die KI systemische Schwachstellen innerhalb einer bestehenden Customer Journey Definition identifizieren, konkrete Optimierungen empfehlen oder schliesslich aktiv bei deren praktischer Umsetzung assistieren kann. Der Erfolg von KI-Initiativen bemisst sich folglich nicht mehr allein an der Reduzierung von Ausgaben. Er definiert sich über eine ausgewogene Kombination aus operationaler Effizienz, inhaltlicher Qualität, messbaren Umsatzauswirkungen sowie kontinuierlichen Verbesserungen der gesamten Customer Journey und der Zufriedenheit der eingesetzten Agenten. ■

DER AUTOR

Josef Novak ist Chief Innovation Officer bei Spitch. Er befasst sich schwerpunktmässig mit der strategischen Entwicklung und Implementierung von dialogorientierten KI-Systemen im Omnichannel-Umfeld. Spitch ist ein global tätiger Anbieter von sprach- und textbasierten KI-Lösungen, der sich unter anderem auf die Anforderungen des Schweizer Marktes sowie die Verarbeitung lokaler Dialekte spezialisiert hat.