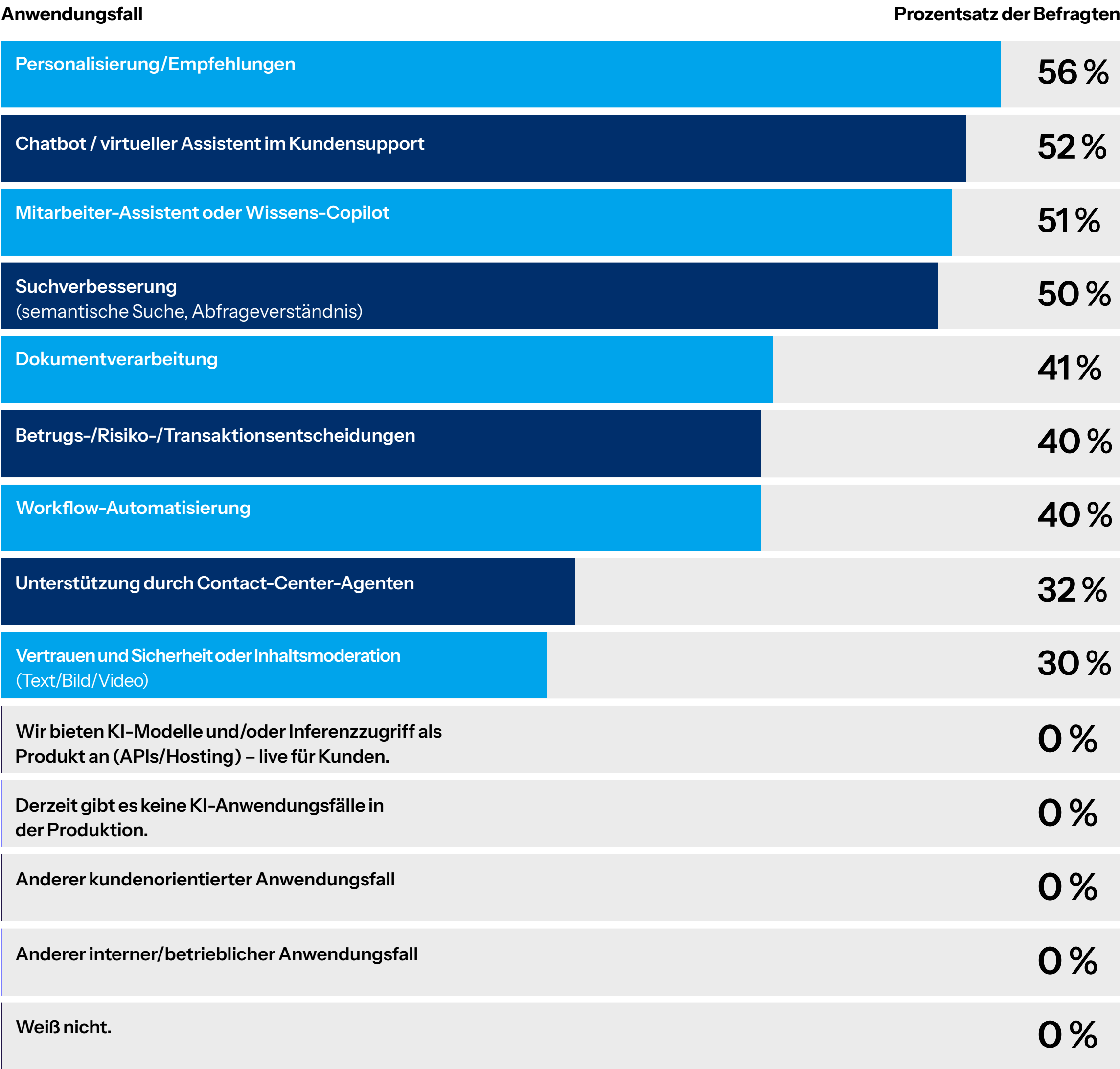


Welche der folgenden KI-Anwendungsfälle setzt Ihr Unternehmen derzeit in der Produktion um?



2. [www.gartner.com](https://www.gartner.com)

Performance steigert die Rentabilität

Wenn die KI-Performance zu wünschen übrig lässt, leidet die Rentabilität in drei wichtigen Bereichen: Die Kundenzufriedenheit wird beeinträchtigt, die Servicekosten steigen und der Umsatz pro Besucher sinkt. Kurz gesagt: Wenn Ihre Inferenz nicht funktioniert, gilt das auch für Ihr Geschäft.

Die Geschäftsmetriken, die nach Angaben der Befragten am stärksten von der Inferenz-Performance betroffen sind, sind unten aufgeführt.

Geschäftsmetriken, die am stärksten von Inferenz-Performance betroffen sind

| Metrik                | Prozentsatz der Befragten |
|-----------------------|---------------------------|
| Kundenzufriedenheit   | 49 %                      |
| Bereitstellungskosten | 40 %                      |
| Umsatz pro Besucher   | 39 %                      |

Wo Geschwindigkeit am wichtigsten ist

Performance ist in allen Bereichen wichtig, doch in Echtzeit-Workflows macht sie den größten Unterschied aus. Wenn KI hinter einer Live-Kundeninteraktion oder einer sekundenschnellen Entscheidung steht, gibt es keinen Spielraum für Fehler. Deshalb verlangen Unternehmen eine schnelle Reaktion bei Personalisierung, bei Entscheidungen zu Betrugsfällen oder Transaktionen, bei Suchen, im Kundensupport sowie in Bezug auf Vertrauen und Sicherheit.

In diesen Situationen, in denen hoher Druck herrscht, ist der Erfolg binär: Entweder ist die Inferenz schnell genug, um die Chance zu nutzen, oder sie ist es nicht. Und wenn sie es nicht ist, gehen sowohl Vertrauen als auch Umsatz verloren.

Dieses Risiko wächst mit zunehmendem Reifegrad des Unternehmens. Im Vergleich zu Unternehmen, die generative KI noch nicht in der Produktion einsetzen, ist es bei Unternehmen mit generativer KI um 45 % wahrscheinlicher, dass sie mit solchen geschwindigkeitskritischen Szenarien konfrontiert sind. Bei Unternehmen mit mehr als 500 Mitarbeitern ist es um 44 % wahrscheinlicher, dass sie latenzempfindliche Workloads ausführen, als bei kleineren Wettbewerbern. Größe bringt nicht nur Komplexität mit sich – sie macht Geschwindigkeit unverzichtbar.

Welche Ihrer KI-Anwendungsfälle in der Produktion erfordern eine schnelle Inferenzreaktion, um gut zu funktionieren?

| Anwendungsfall                                  | Prozentsatz der Befragten |
|-------------------------------------------------|---------------------------|
| Personalisierung                                | 54 %                      |
| Betrug, Risiko oder Transaktionen               | 49 %                      |
| Suchen                                          | 44 %                      |
| Kundensupport                                   | 41 %                      |
| Vertrauen und Sicherheit oder Inhaltsmoderation | 40 %                      |

Der Großteil der KI-Ausgaben im Jahr 2026 wird auf Inferenz-Workloads entfallen.<sup>2</sup>