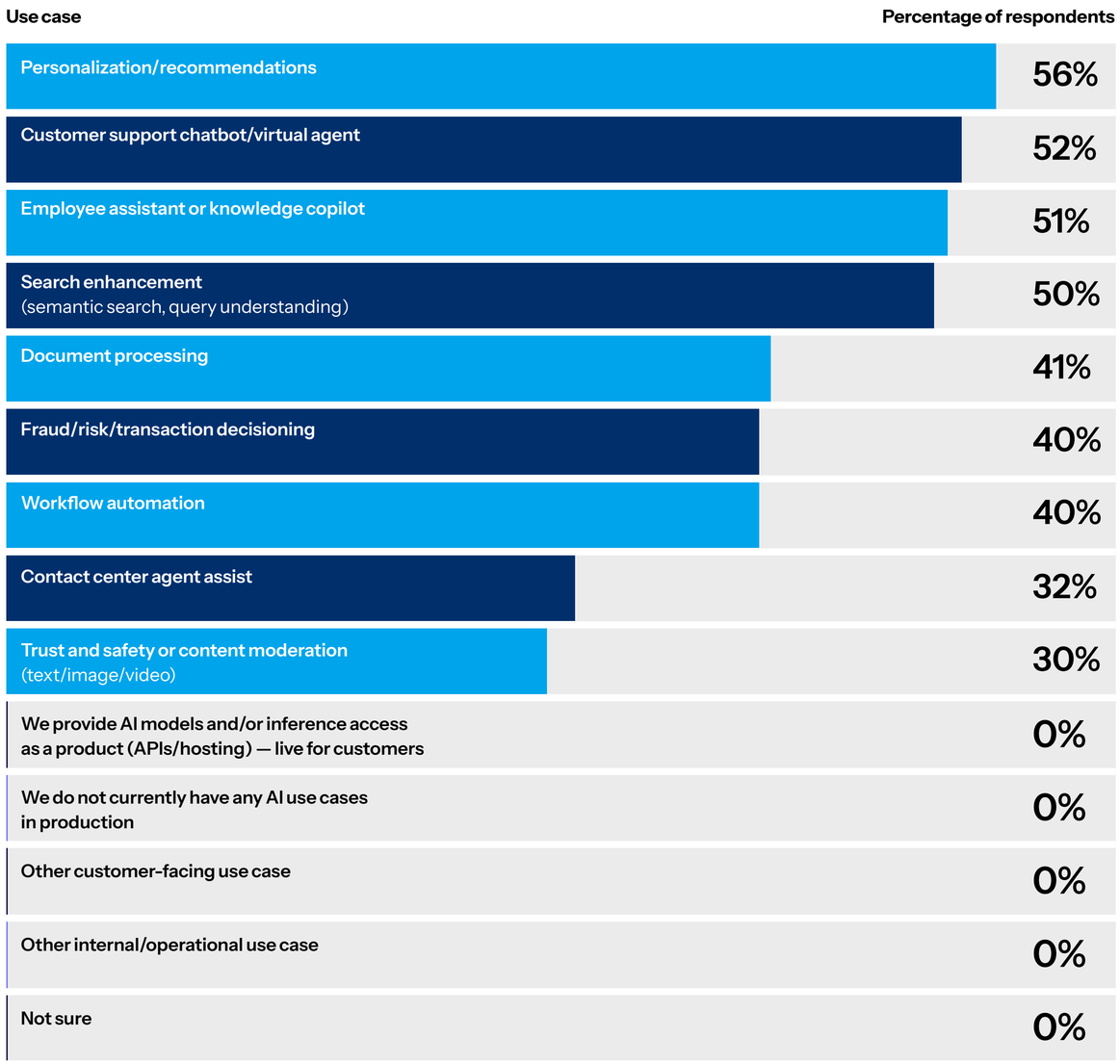
Which of the following AI use cases are currently in production at your organization?



Performance drives profitability

When AI performance lags, profitability suffers across three critical fronts: it erodes customer satisfaction, inflates cost to serve, and suppresses revenue per visitor. In short, if your inference isn't performing, your business isn't either.

The business metrics most affected by inference performance, according to respondents, are shown below.

Business metrics most affected by inference performance

Metric	Percentage of respondents
Customer satisfaction	49%
Cost to serve	40%
Revenue per visitor	39%

Where speed matters most

Performance matters across the board, but it moves the needle most aggressively in real-time workflows. When AI sits behind a live customer interaction or a split-second decision, the margin for error vanishes. This is why organizations demand rapid responses for personalization, fraud or transaction decisioning, search, customer support, and trust and safety.

In these high-pressure moments, success is binary. Inference is either fast enough to capture the opportunity or it isn't. And if it's not, both trust and revenue will be lost.

This risk only grows as companies mature. Compared with organizations that have not yet deployed GenAI in production, businesses with GenAI already in production are 45% more likely to face these speed-critical scenarios. Organizations with 500+ employees are 44% more likely to run latency-sensitive workloads than smaller competitors. Scale does not just bring complexity — it makes speed essential.

Of the AI use cases you have in production, which require a rapid response from inference to work well?

Use case	Percentage of respondents
Personalization	54%
Fraud, risk, or transaction	49%
Search	44%
Customer support	41%
Trust and safety, or content moderation	40%

> **The majority of AI spending in 2026 will be on inference workloads**[2]

2. www.gartner.com