

Personal Agents sicher testen

Was persönliche KI-Agenten können und wie du die Rechte im Test begrenzt

Ein **Personal Agent** ist ein KI-System, das einen Auftrag über einen längeren Zeitraum verfolgt: Er recherchiert, bedient Programme und meldet sich **von selbst**, wenn er ein Ergebnis liefert oder eine Freigabe braucht. Ob er seine Arbeit gemacht hat, zeigt ein **Ergebnis, das du nachprüfen kannst**: ein Kalendereintrag, eine Buchungsnummer, eine Gutschrift.

CHATBOT ODER AGENT: ERKLÄREN GEGEN AUSFÜHREN

Chatbot	Personal Agent
Beantwortet direkte Fragen.	Verfolgt einen Auftrag im Hintergrund weiter.
Erklärt, wie eine Buchung funktioniert.	Führt die Buchung selbstständig durch.
Liefert einen Text.	Liefert eine fertige Buchung oder einen Kalendereintrag.

GEEIGNET: WARTEN, NACHHAKEN, VERGLEICHEN

- **Messetermin beobachten:** Agent trägt den Termin ein, sobald er feststeht. Beleg ist der korrekte Kalendereintrag.
- **Tarife vergleichen:** Agent schlägt einen Wechsel vor. Du prüfst Zahlen und Konditionen, bevor du den Vertrag freigibst.
- **Gutschrift nachfassen:** Erledigt erst, wenn das Geld tatsächlich auf dem Konto eingeht.

DER TESTPLAN: EINE WOCHE, EIN AUFTRAG

- 1 **Mit einem Testkonto arbeiten**, das Hauptpostfach bleibt draußen. Nur die Informationen geben, die dieser eine Auftrag braucht.
- 2 **Rechte auf Lesezugriff und Entwürfe.** Käufe, Kündigungen und den Mailversand immer erst freigeben lassen.
- 3 **Zahlungen nur mit Einmalkarte** oder einer Kreditkarte mit sehr niedrigem Limit.
- 4 **Protokoll regelmäßig prüfen**, spätestens nach sieben Tagen. Jeden gemeldeten Erfolg an echten Belegen kontrollieren.
- 5 **Handelt der Agent unerwartet, sofort abbrechen.** Rechte erst erweitern, wenn er einen klaren Nutzen beweist.

VIER RISIKEN

- **Prompt Injection:** Versteckte Anweisungen in Webseiten, Dokumenten oder Mails lenken den Agenten um. Laut Meta branchenweit ungelöst.
- **Zu viele Rechte:** Postfach, Dateien und Bankkonto zugleich: Ein Fehler kann großen Schaden anrichten. Bei dots (OpenAI) löscht das Trennen einer App abgerufene Informationen nicht.
- **Nicht umkehrbar:** Aktionen lassen sich nicht immer rückgängig machen.
- **Daten:** Training mit Nutzerdaten ist bei Muse (Meta) und Instinct abschaltbar, bei Instinct mit Ausnahmen für Sicherheitsprüfungen.

DER PROMPT AUS DEM BEITRAG: VERTRÄGE PRÜFEN MIT FREIGABE-GRENZE

ZUM KOPIEREN

„Ich möchte meine laufenden Verträge für Handy, Internet und Strom überprüfen. Du bekommst dafür nur Lesezugriff auf die Unterlagen, die ich dir in diesem Test freigebe. Erstelle eine Tabelle mit Anbieter, monatlichen Kosten, Kündigungsfrist und Vertragsende. Suche für jeden Vertrag ein aktuelles Vergleichsangebot und rechne den Unterschied auf ein Jahr hoch. Wenn dir Angaben fehlen, frag nach, statt zu schätzen. Du darfst recherchieren und Entwürfe für Kündigungsschreiben vorbereiten. Kündigen, abschließen, kaufen oder Nachrichten in meinem Namen senden darfst du nur nach meiner ausdrücklichen Freigabe.“

FÜR BETRIEBE IN ÖSTERREICH

KUNDENDATEN

Der Anbieter ist in der Regel Auftragsverarbeiter: Vertrag nach Art. 28 Abs. 3 DSGVO nötig, ein privates Konto reicht in der Regel nicht.

DATENTRANSFER

Fließen Daten in ein Drittland, gelten die Regeln zum internationalen Datentransfer.

KI-KOMPETENZ

Art. 4 AI Act seit 02.02.2025: Das Team sollte wissen, wann es prüft und wann es abbricht.

MERKSÄTZE

Erledigt ist, was du nachprüfen kannst.

Erst wenig Rechte, mehr erst nach bewiesenem Nutzen.