CHECKLISTE

RAG- oder KI-Assistenten-Projekt richtig scopen: eine Checkliste

30 Fragen, die vor dem Bau eines RAG-Systems oder KI-Assistenten geklärt sein sollten: Use Case, Daten, Zugriffe, Retrieval, Modellwahl, Evaluation und Betrieb.

30 Prüfpunkte	Haken Sie jeden Punkt ab, sobald er umgesetzt ist. Die Online-Version enthält dieselben Erläuterungen mit Links.

Online-Version: sigmacode.io/de/insights/rag-ai-assistant-scoping-checklist

Die meisten KI-Assistenten-Projekte, die enttäuschen, scheitern nicht am Modell. Sie scheitern daran, dass nie geklärt wurde, wofür der Assistent eigentlich da ist, dass die zugrunde liegenden Inhalte lückenhaft oder veraltet waren, dass Berechtigungen erst spät bedacht wurden oder dass niemand messen konnte, ob eine Änderung etwas verbessert oder verschlechtert hat. Die Modellwahl ist wichtig, aber meist eine der leichteren Entscheidungen.

Diese Checkliste fasst die Fragen zusammen, die wir klären, bevor wir ein Retrieval-Augmented-Generation-System (RAG) oder einen KI-Assistenten auf Unternehmensdaten bauen. Nutzen Sie sie, um ein internes Projekt vorzubereiten, einen Dienstleister zu briefen oder zu prüfen, ob ein vorliegendes Angebot das Wesentliche abdeckt. Nicht jede Antwort muss am ersten Tag feststehen, aber Sie sollten wissen, welche noch offen sind.

1. Use Case und Erfolgskriterien

Halten Sie diese Punkte schriftlich fest, bevor jemand ein Notebook öffnet oder Modelle vergleicht.

- [ ] 01 **Eine Hauptaufgabe.** Benennen Sie die eine Aufgabe, die der Assistent zuerst gut erledigen muss, zum Beispiel Produktfragen des Support-Teams beantworten oder Klauseln in Lieferantenverträgen finden. Weitere Use Cases können folgen, sobald der erste gemessen ist; mehrere zugleich zu starten verwässert Anforderungen und Evaluation.
- [ ] 02 **Nutzer und ihr Kontext.** Beschreiben Sie, wer fragt, wie oft, in welcher Sprache und auf welchem Gerät, und was mit der Antwort geschieht. Eine interne Fachkraft, die Quellen prüft, braucht etwas anderes als ein Kunde, der direkt auf Basis der Antwort handelt.
- [ ] 03 **Echte Fragen mit idealen Antworten.** Sammeln Sie 20–50 echte Fragen aus Tickets, E-Mails oder Chatverläufen und lassen Sie eine Fachexpertin oder einen Fachexperten die ideale Antwort samt passender Quelle formulieren. Dieses Set definiert, was „gut“ bedeutet, und bildet den Grundstock Ihres Evaluationssets.
- [ ] 04 **Abgrenzung und Ablehnungen.** Legen Sie fest, was der Assistent nicht tun darf: Rechts-, Medizin- oder Finanzberatung geben, über seine Quellen hinaus antworten, im Namen des Unternehmens Zusagen machen. Bestimmen Sie auch, was er stattdessen sagt und wohin er verweist.

- [ ] 05 **Baseline des heutigen Prozesses.** Dokumentieren Sie, wie die Aufgabe heute erledigt wird, wie lange sie dauert, wer sie übernimmt und was sie kostet. Ohne Baseline ist „der Assistent hilft“ eine Meinung, kein Ergebnis.

2. Datenquellen

Die Qualität der Antworten ist durch die Qualität und Zugänglichkeit der dahinterliegenden Inhalte begrenzt.

- [ ] 06 **Quelleninventar.** Listen Sie jede Quelle auf, die der Assistent nutzen soll: Wikis, SharePoint- oder Google-Drive-Ordner, Ticketsysteme, Datenbanken, PDFs, Websites. Notieren Sie jeweils, wie der Zugriff erfolgt (API, Export, Crawling) und ob er zulässig ist.
- [ ] 07 **Formate und Dokumentqualität.** Prüfen Sie, ob es gescannte PDFs gibt, die OCR benötigen, sowie komplexe Tabellen, Formulare, Folien und Bilder mit wesentlichem Inhalt. Tabellen und Scans sind die Stellen, an denen einfache Pipelines am meisten Information verlieren, daher sollten Sie sie früh stichprobenartig prüfen.
- [ ] 08 **Verantwortung und Aktualität.** Benennen Sie für jede Quelle eine verantwortliche Person und halten Sie fest, wie oft sie sich ändert. Wenn niemand für einen Inhalt zuständig ist, korrigiert auch niemand die falschen Antworten, die daraus entstehen.
- [ ] 09 **Dubletten, Versionen und Sprachen.** Identifizieren Sie veraltete Kopien, Entwürfe und parallele Fassungen derselben Richtlinie und legen Sie fest, welche maßgeblich ist. Erfassen Sie die Sprachen von Dokumenten und Fragen, denn sprachübergreifendes Retrieval muss ausdrücklich getestet werden.

3. Zugriffskontrolle und Datenschutz

Klären Sie diese Punkte, bevor echte Daten Ihre Systeme verlassen.

- [ ] 10 **Berechtigungen im Retrieval.** Wenn Nutzer nur bestimmte Dokumente sehen dürfen, muss das Retrieval nach den Berechtigungen der anfragenden Person filtern, synchronisiert aus den Quellsystemen. Anweisungen im Prompt oder das Vertrauen darauf, dass das Modell Inhalte zurückhält, sind keine Zugriffskontrolle.
- [ ] 11 **Personenbezogene Daten und Rechtsgrundlage.** Ermitteln Sie personenbezogene Daten in Dokumenten, Fragen und Logs und dokumentieren Sie Rechtsgrundlage und Zweck der Verarbeitung nach DSGVO. Beziehen Sie Ihre Datenschutzbeauftragten zu Beginn ein, nicht erst zum Launch.
- [ ] 12 **Anbietervertrag und Datenresidenz.** Schließen Sie mit jedem Modell- und Infrastrukturanbieter einen Auftragsverarbeitungsvertrag (AVV), prüfen Sie die Unterauftragsverarbeiter und stellen Sie sicher, dass die Verarbeitung bei Bedarf in einer EU-Region bleibt. Lassen Sie sich schriftlich bestätigen, dass Ihre Daten nicht zum Training der Modelle des Anbieters verwendet werden.
- [ ] 13 **Aufbewahrung, Logging und Schwärzung.** Legen Sie fest, wie lange Prompts, abgerufene Passagen und Antworten gespeichert werden, wer sie lesen darf und ob personenbezogene Daten vor dem Logging geschwärzt werden. Logs werden für Fehlersuche und Evaluation gebraucht; das Ziel ist also eine kontrollierte Aufbewahrung, nicht der Verzicht auf Logs.

4. Retrieval-Design

Die meisten falschen Antworten in RAG-Systemen gehen auf das Retrieval zurück, nicht auf das Modell; bei einem kleinen, stabilen Bestand kann Long Context mit Prompt Caching das Retrieval ganz ersetzen (siehe RAG oder Fine-Tuning).

- [ ] 14 **Chunking entlang der Dokumentstruktur.** Teilen Sie Inhalte an Überschriften, Abschnitten, Listenpunkten und Tabellengrenzen auf statt nach fester Zeichenzahl. Geben Sie jedem Chunk den Überschriftenpfad mit, damit eine Passage auch für sich allein verständlich bleibt.
- [ ] 15 **Metadaten und Filter.** Speichern Sie zu jedem Chunk Titel, Quelle, Abschnitt, Datum, Sprache, Version und Berechtigungsgruppen. Erst Metadaten ermöglichen Berechtigungsfilter, Regeln wie „nur die aktuelle Version“ und brauchbare Quellenangaben.
- [ ] 16 **Hybride Suche und Reranking.** Kombinieren Sie Keyword-Suche für exakte Begriffe wie Produktcodes, Artikelnummern und Namen mit Vektorsuche über Embeddings für die Bedeutung, und ordnen Sie die gemeinsamen Kandidaten anschließend per Reranking neu. Testen Sie jeden Schritt mit Ihren Beispielfragen, statt anzunehmen, dass die Standardeinstellungen genügen.
- [ ] 17 **Quellenangaben als Pflicht.** Verlangen Sie, dass der Assistent die verwendeten Passagen zitiert, mit Link auf Quelldokument und Abschnitt. Quellenangaben erlauben Nutzern die Prüfung und zeigen Reviewern, ob eine falsche Antwort aus schlechtem Retrieval oder schlechter Generierung stammt.

5. Modell- und Anbieterwahl

Wählen Sie das Modell erst, wenn ein Evaluationsset vorliegt, damit die Entscheidung auf Ihren Daten beruht und nicht auf allgemeinen Benchmarks.

- [ ] 18 **Hosting-Variante.** Vergleichen Sie eine gehostete API, dieselben oder ähnliche Modelle in einer EU-Cloud-Region und ein selbst betriebenes Open-Weight-Modell auf eigener Infrastruktur. Jede Variante verschiebt die Balance zwischen Antwortqualität, Datenhoheit, Betriebsaufwand und Kosten.
- [ ] 19 **Latenz und Kosten pro Anfrage.** Messen Sie die Antwortzeit von Ende zu Ende und die Kosten pro beantworteter Frage anhand Ihrer eigenen Beispielfragen, einschließlich Retrieval, Reranking sowie Input- und Output-Tokens. Prompt Caching und kleinere Modelle für einfache Teilschritte können diese Werte deutlich verändern.
- [ ] 20 **Fallback und Lock-in.** Planen Sie, was passiert, wenn der Anbieter ausfällt oder ein Modell abkündigt: ein zweiter Anbieter, ein eingeschränkter Modus oder eine klare Fehlermeldung. Halten Sie Prompts, Evaluationsset und Retrieval-Schicht anbieterneutral, sodass ein Wechsel eine Konfigurationsänderung plus Eval-Lauf ist und keine Neuentwicklung.

6. Evaluation

Wer Qualität nicht messen kann, kann sie weder verbessern noch die Go-live-Entscheidung begründen.

- [ ] 21 **Evaluationsset vor dem Bau.** Machen Sie aus den echten Fragen aus Abschnitt 1 ein versioniertes Evaluationsset mit erwarteten Antworten und erwarteten Quellen und ergänzen Sie es um schwierige Fälle und Grenzfälle. Bauen Sie es vor dem ersten Prototyp auf, nicht nach der ersten Demo.

- [ ] 22 **Retrieval- und Antwortqualität.** Messen Sie getrennt, ob die richtigen Passagen gefunden wurden und ob die Antwort korrekt, vollständig und diesen Passagen treu ist. Prüfen Sie die Genauigkeit der Quellenangaben: Die zitierte Passage muss die Aussage tatsächlich stützen.
- [ ] 23 **Tests für Ablehnungen und Prompt Injection.** Nehmen Sie Fragen auf, die der Assistent ablehnen muss, und Fragen, die seine Quellen nicht beantworten, und prüfen Sie, dass er dies offen sagt, statt zu raten. Ergänzen Sie Dokumente und Eingaben, die seine Anweisungen aushebeln oder Daten anderer Nutzer abgreifen sollen, und stellen Sie sicher, dass sie scheitern.
- [ ] 24 **Regressionsläufe und menschliche Prüfung.** Führen Sie das komplette Evaluationsset bei jeder Änderung an Prompts, Modellen, Chunking oder Daten aus und vergleichen Sie mit dem vorherigen Lauf. Automatische Bewertung durch ein Modell beschleunigt das, doch eine Fachperson sollte die Ergebnisse regelmäßig stichprobenartig prüfen.

7. Integration und Betrieb

Der Assistent braucht einen festen Ort, und jemand muss ihn nach dem Launch betreiben.

- [ ] 25 **Kanal und Anmeldung.** Entscheiden Sie, wo der Assistent eingesetzt wird: als Widget auf der Website, in Slack oder Microsoft Teams, in einem internen Tool oder in einem bestehenden Ticketsystem. Nutzen Sie Ihr vorhandenes SSO, damit Identität und Berechtigungen aus derselben Quelle kommen wie überall sonst.
- [ ] 26 **Übergabe an Menschen und Feedback.** Legen Sie fest, wie Nutzer eine Person erreichen, wenn der Assistent nicht weiterhilft, und zwar mit dem bisherigen Gesprächsverlauf. Fügen Sie einfache Feedback-Buttons hinzu und führen Sie negatives Feedback in das Evaluationsset zurück.
- [ ] 27 **Kostenmodell und Monitoring.** Schätzen Sie die Kosten pro Anfrage und pro Monat bei erwarteter und bei Spitzenlast und richten Sie Budgetwarnungen ein. Überwachen Sie Latenz, Fehlerquoten, Ablehnungsquoten und Nutzerfeedback, mit Logs, die wie in Abschnitt 3 vereinbart geschwärzt sind.
- [ ] 28 **Neuindexierung, Inhaltsverantwortung und Incidents.** Automatisieren Sie die Neuindexierung bei Änderungen an Quelldokumenten, einschließlich Löschungen, damit der Assistent nie zurückgezogene Inhalte zitiert. Benennen Sie, wer Inhaltsprobleme behebt, und definieren Sie einen Incident-Prozess für falsche, schädliche oder unbefugt offengelegte Antworten, einschließlich einer schnellen Abschaltmöglichkeit.

8. Go/No-Go-Kriterien

Vereinbaren Sie die Entscheidungsregeln, bevor Ergebnisse vorliegen, damit nicht eine gelungene Demo die Entscheidung trifft.

- [ ] 29 **Vorab vereinbarte Schwellenwerte.** Halten Sie Mindestwerte für Antwortkorrektheit und Genauigkeit der Quellenangaben im Evaluationsset fest, dazu akzeptable Latenz und Kosten pro Anfrage sowie null Toleranz für Antworten, die Berechtigungsgrenzen überschreiten. Vergleichen Sie die Ergebnisse mit der Baseline aus Abschnitt 1.
- [ ] 30 **Zeitlich begrenzter Pilot mit klarem Ausstieg.** Führen Sie einen Pilot mit einer kleinen Gruppe echter Nutzer über einen festen Zeitraum durch, mit einer benannten Person, die entscheidet. Den Umfang zu verkleinern oder abzubrechen ist ein legitimes Ergebnis und weit günstiger, als es erst nach einem vollständigen Rollout festzustellen.

Wie wir helfen können

Wenn Sie diese Checkliste gemeinsam mit uns durcharbeiten möchten, ist unser **AI Proof-of-Value Sprint** mit festem Umfang der direkteste Weg. In zwei Wochen schärfen wir mit Ihnen den Use Case, bauen einen RAG- oder Agenten-Prototyp auf Ihren eigenen Daten, erstellen ein Evaluationsset, erarbeiten ein Kostenmodell und liefern einen Go/No-Go-Bericht, den Sie Ihren Entscheidern vorlegen können. Details finden Sie auf unserer Seite zu Preisen und bei unseren Leistungen für KI & Automatisierung.

Bevor eines Ihrer Dokumente ein Modell erreicht, vereinbaren wir mit Ihnen, wie mit den Daten umgegangen wird; vorab können Sie nachlesen, wie wir mit Kundendaten in KI-Projekten umgehen. Wenn Sie lieber mit einem Gespräch beginnen, fordern Sie eine kostenlose Einschätzung an und beschreiben Sie Ihren geplanten Use Case.