cloumo.

Maßgeschneiderter KI-Recherche-Agent für Projektunterlagen mit je über 1.000+ Dokumenten

Warum klassische RAG-Systeme scheitern und wie Multimodalität & Tool Calling den Unterschied machen

Technologie	Architektur
LangChain · Qdrant · Claude	Agent-based Retrieval-Augmented Generation (RAG)

Ergebnis

> 90% Zeitersparnis bei der Dokumentenrecherche

cloumo.

Steckbrief

Das Problem

Ein mittelständisches Unternehmen stand vor einem klassischen Wissensproblem: bis zu **1.000+ Dokumente je Projekt** – darunter Verträge, Projektunterlagen, Richtlinien – waren vorhanden, aber praktisch unzugänglich. Manuelle Recherchen kosteten Stunden, kritische Details wurden übersehen, und Standard-KI-Lösungen lieferten unzuverlässige Antworten ohne Quellennachweis.

Die Lösung

Mit einer maßgeschneiderten **Agent-basierten RAG-Architektur** auf Basis von Vision LLM, LangChain, Qdrant und Claude (Anthropic) wurde ein interner KI-Recherche-Agent entwickelt, der komplexe Fragen in 1 bis 2 Minuten beantwortet – präzise und nachvollziehbar.

> 90%

Zeitersparnis

von durchschnittlich 45 Min. auf 2 Min. reduziert

100%

Nachvollziehbarkeit

Grounded Generation auf Basis eigener Dokumente

cloumo.

Die Herausforderung

Ausgangslage

1.000+ Dokumente je Projekt

Verträge, Projektunterlagen, interne Richtlinien – verteilt, unstrukturiert, schwer durchsuchbar

Manuelle Suche

Komplexe Recherchen dauerten **30–60 Minuten** pro Anfrage – ein massiver Produktivitätsverlust

Übersehene Details

Mitarbeiter konnten nicht sicherstellen, alle relevanten Passagen gefunden zu haben

Pain Points

Halluzinationen

Standard-KI-Lösungen erfanden Inhalte – ein absolutes No-Go im Unternehmenskontext

Keine Nachvollziehbarkeit

Antworten ohne Quellenangabe sind nicht vertrauenswürdig und nicht revisionssicher

Skalierbarkeit

Mit wachsendem Dokumentenbestand wurde stieg der Aufwand stetig

cloumo.

Warum klassische RAG-Systeme scheitern

und wie **Multimodalität** & **Tool Calling** den Unterschied machen

Klassisches RAG

Starres Chunking

Dokumente werden blind in gleich große Stücke zerlegt, Kontext geht verloren

Kein Verständnis für Tabellen & Bilder

Nur Text wird verarbeitet, multimodale Inhalte werden ignoriert

Single-Shot Retrieval

Eine einzige Suchanfrage, egal wie komplex die Frage ist

Keine Selbstkorrektur

Das System merkt nicht, wenn die gefundenen Passagen irrelevant sind

Agentic RAG – Der Unterschied

Intelligentes Chunking

Semantisch sinnvolle Segmentierung, Kontext bleibt erhalten

Multimodalität

Text, Tabellen, Grafiken und Bilder werden gemeinsam verstanden

Iteratives Tool Calling

Der Agent stellt mehrere gezielte Suchanfragen bis die Antwort vollständig ist

Selbstreflexion

Claude bewertet die Qualität der Ergebnisse und korrigiert den Kurs eigenständig

cloumo.

Die Lösung: **Agentic** RAG-Architektur

Warum klassische RAG-Systeme scheitern und wie Multimodalität & Tool Calling den Unterschied machen

Statt einer starren Pipeline agiert das System als **autonomer Recherche-Agent**. Über **Advanced Tool Calling** entscheidet das Modell selbstständig, welche Suchwerkzeuge es wie oft einsetzt. Findet der Agent in einem Dokument einen Verweis auf einen anderen Vertrag, ruft er dieses eigenständig als nächstes ab, um die verknüpfte Information zu prüfen – genau wie ein menschlicher Analyst.

Nutzerfrage

LangChain Orchestriert Agent

Agent entwickelt Strategie

Datenbank durchsuchen

Vollständige Antwort synthetisieren

LangChain

Orchestrator: Automatisiertes Chunking, Indexierung & Steuerung der Abfragelogik

Qdrant

Vektordatenbank: Blitzschnelle semantische Suche über den gesamten Dokumentenbestand

Claude (Anthropic)

Großes Kontextfenster, präzise Textanalyse, nachweislich geringe Halluzinationsrate

Grounded Generation

Antworten ausschließlich auf Basis der eigenen Dokumente – keine Erfindungen

cloumo.

Das Quellen-Feature: Vertrauen durch Transparenz

Das Herzstück der Lösung ist die **lückenlose Rückverfolgbarkeit** jeder Antwort – von der Frage bis zur exakten Fundstelle im Originaldokument.

01

Frage stellen

Nutzer stellt eine Frage im UI – z. B. *„Welche Kündigungsfristen gelten in Vertrag X?"*

02

Semantische Suche

LangChain orchestriert die Agenten und diese suchen via Qdrant die relevantesten Textsegmente aus den 1.000 Dokumenten

03

Antwortgenerierung

Die Segmente + Frage werden an Claude übergeben – Claude antwortet **ausschließlich** auf Basis dieser Passagen

04

Quellenverknüpfung

Segment-IDs werden mitgeführt: Jede Aussage ist mit Dateiname & Seitenzahl verknüpft

05

Verifikation per Klick

Nutzer sehen die Antwort **und** können per Klick die exakte Quelle öffnen

Beispiel-Output im UI

Frage: *„Welche Kündigungsfristen gelten in Vertrag X?"*

KI-Antwort: „Gemäß Abschnitt 8.2 gilt eine Kündigungsfrist von 3 Monaten zum Quartalsende..."

[Quelle: Vertrag_Firma_B.pdf, Seite 14]

Dateiname

Jede Antwort zeigt das Quelldokument

Seitenzahl

Exakte Fundstelle auf Seitenebene

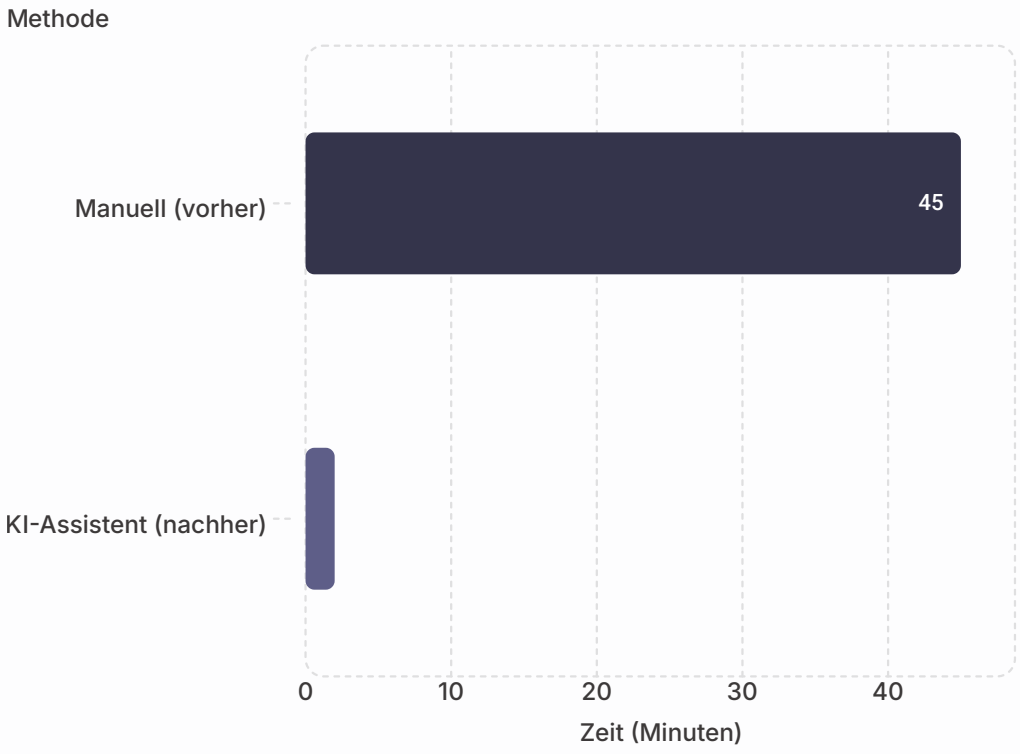
Direktlink

Ein Klick öffnet das Original

cloumo.

Ergebnisse & ROI

Vorher / Nachher: Recherchezeit



Dramatische Zeitersparnis: Von **45 Minuten** manueller Suche auf **2 Minuten** mit dem KI-Assistenten – eine Reduktion um bis zu **90 %**.

Messbare Erfolge

> 90%

Zeitersparnis

45 Min → 2 Min pro Recherche

100%

Quellennachweis

Jede Antwort mit verifizierbarer Quellenangabe

Technologie-Stack auf einen Blick

Komponente	Rolle	Vorteil
LangChain	Orchestrator	Flexibles Chunking & Retrieval
Qdrant	Vektordatenbank	Schnelle semantische Suche
Claude	LLM / Analyst	Großes Kontextfenster, geringe Halluzinationsrate
RAG	Architekturmuster	Grounded Generation, keine Erfindungen

Open-Source-Kern

LangChain & Qdrant sind battle-tested, community-getrieben und produktionsreif

Enterprise-LLM

Claude (Anthropic) mit Enterprise Agreement für Daten Souveränität, großem Kontextfenster und Best-In-Class Modell für Agenten.

Custom UI

Maßgeschneidertes Interface, gemäß Corporate Design.

cloumo.

Ihr nächster Schritt: Maßgeschneiderte KI für Ihr Unternehmen

Was wir für Sie tun können

Intelligente Assistenzsysteme & RAG

Dokumenten- und Wissensmanagement der nächsten Generation.

Prozessautomatisierung (Agents)

Autonome KI-Agenten, die wiederkehrende Workflows und administrative Aufgaben für Sie übernehmen.

Custom LLM- & Vision-Integration

Anbindung und Feintuning modernster Sprach- und Bilderkennungsmodelle für Ihre Kernprozesse.

KI-Strategie & Prototyping

Von der ersten Potenzialanalyse bis zum einsatzbereiten Proof of Concept in wenigen Wochen.

Lassen Sie uns über Ihre Vision sprechen

Egal ob Sie einen riesigen Dokumentenbestand bändigen, Workflows automatisieren oder eine völlig neue KI-Idee umsetzen möchten – wir zeigen Ihnen im kostenlosen 30-minütigen Demo-Call, was machbar ist.

30-Min. Demo-Call

Live-Demonstration mit einem realen Dokumentenbestand – sehen Sie die Technologie in Aktion.

Individuelle Analyse

Wir analysieren Ihren spezifischen Use Case und zeigen den konkreten ROI-Potenzial auf.

Schneller Einstieg

Vom ersten Gespräch zum laufenden PoC in weniger als zwei Monaten.

Bereit für den nächsten Schritt? Schreiben Sie uns: info@cloumo.com