Private AI Agent Lösung: Knowledge Base, Chat und Workflow-Automation

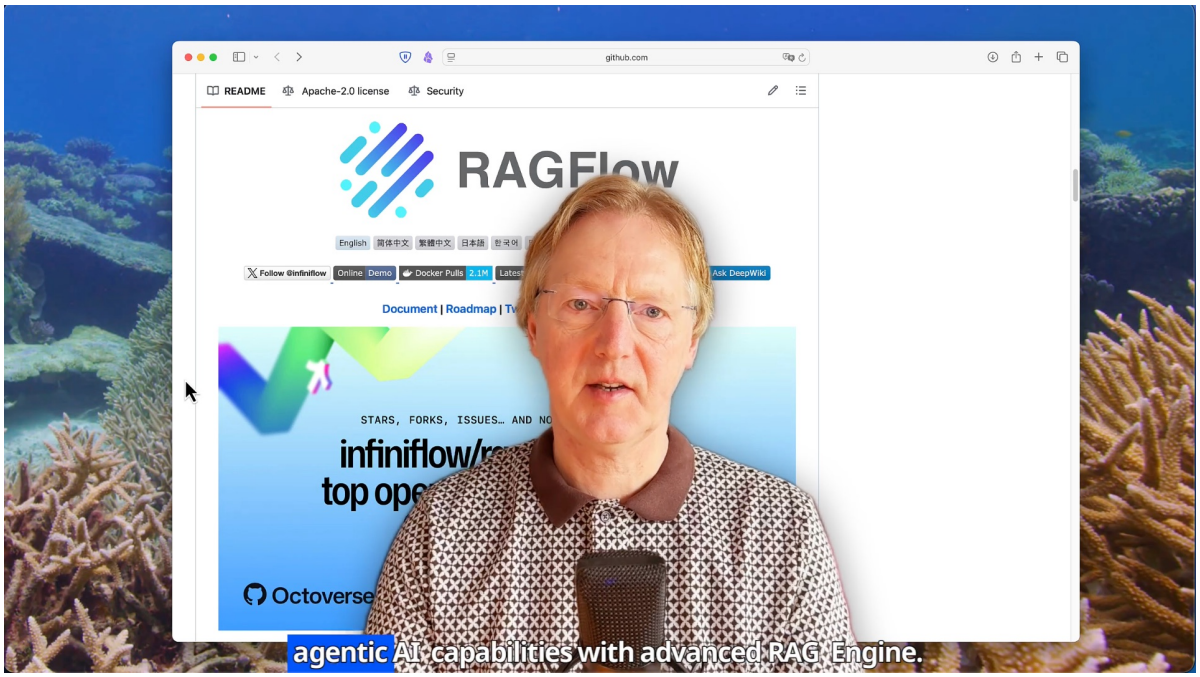
Andreas Wittmann

andreas.wittmann@mailbox.org

Datum: 2025-11-11

Die Private AI Agent Solution ist eine vollständig konfigurierte RagFlow-Installation, die Knowledge Base, Chat-Client und KI-Agenten für die intelligente Verarbeitung Ihrer Unternehmensdokumente vereint - wahlweise mit vertrauenswürdigen Cloud-APIs (Standard, 699) oder vollständig Air-gapped mit eigenen Open-Source-Modellen (Professional, 1200/Tag).
Download: PDF-Version dieses Dokuments.

1 Einführung



Dieses Video zeigt die Bereitstellung von RAGFlow auf AWS unter Verwendung des Live-Scripting-Ansatzes. Die Bereitstellung nutzt das Open-Source-Projekt **private-rag** das Basis für die Angebote **Starter** und **Standard** ist. `https://www.youtube.com/watch?v=fLg-f9MkRMo`

Die *Private AI Agent Lösung* ist eine vollständig konfigurierte Installation des Open-Source Frameworks *RagFlow*. Diese Lösung vereint Knowledge Base, Chat-Client und KI-Agenten zu einer integrierten Plattform für die intelligente Verarbeitung Ihrer Unternehmensdokumente.

RagFlow verarbeitet diverse Dokumentenformate durch intelligentes Parsing und semantische Indexierung. Das Ergebnis sind präzise Antworten mit direkten Quellenverweisen, die nachvollziehbare und halluzinationsfreie Kommunikation gewährleisten. Die Lösung ermöglicht Ihren Teams, Unternehmenswissen systematisch zu erschließen - von interaktiven Frage-Antwort-Dialogen bis zu automatisierten Workflows.

Entwickelt für Pilotgruppen von 5-30 Benutzern, nutzt die Installation mein erprobtes private-rag Projekt. Vier Bereitstellungsoptionen stehen zur Verfügung: AWS, Azure, Google Cloud oder On-Premise.

2 Drei Angebotstypen

Figure 1: Übersicht der Private AI Agent Lösungsangebote

2.1 Starter (Kostenlos)

Technisch versierte Teams erhalten über mein GitHub Repository Zugang zur vollständigen Deployment-Dokumentation. Die schrittweise Anleitung ermöglicht eigenständige Technologie-Evaluation vor einer

kommerziellen Implementierung.

2.2 Standard (699,- pro Installation)

Das Standard-Paket liefert eine vollständig konfigurierte RagFlow-Installation mit Knowledge Base, Chat-Client und Agent-Builder. Die Installation erfolgt an einem Tag nach einer Vorbereitungsphase und umfasst Docker-containerisierte Instanz, SSL/HTTPS-gesicherten Zugriff, Terraform-basiertes Setup und vollständige Dokumentation.

Diese Konfiguration nutzt externe KI-APIs für Sprachmodelle und Embeddings. Informationen werden dabei an die von Ihnen gewählten KI-Anbieter weitergeleitet. Sie wählen vertrauenswürdige Anbieter wie AWS Bedrock, Azure OpenAI oder Google Vertex AI basierend auf Ihrer bestehenden Cloud-Infrastruktur.

2.3 Professional (1200,- pro Tag)

Das Professional-Paket ermöglicht den Einsatz eigener Open-Source-Modelle für vollständige Datenkontrolle. Diese Installation kann als Air-gapped System ohne externe API-Abhängigkeiten betrieben werden - alle Komponenten operieren ausschließlich in Ihrer Infrastruktur.

Der Aufwand für diese Konfiguration ist deutlich höher, da eigene Embedding-Modelle und Sprachmodelle bereitgestellt werden müssen. Typische weitere Szenarien umfassen Integration in Authentifizierungssysteme, Anpassung der Dokumenten-Pipeline, Performance-Optimierung oder Migration bestehender Wissensdatenbanken.

3 Kernfunktionen

Knowledge Base: Intelligentes Parsing verarbeitet PDFs, Word, Excel, PowerPoint und weitere Formate. Template-basiertes Chunking erhält semantische Zusammenhänge. Sie kontrollieren die Verarbeitung durch Visualisierung der Text-Chunks und können bei Bedarf manuell eingreifen.

Chat-Client: Natürlichsprachliche Fragen werden mit präzisen Quellenverweisen beantwortet. Spezialisierte Assistenten können für verschiedene Fachbereiche konfiguriert werden, jeweils mit eigener Knowledge Base.

KI-Agenten: Der No-Code Workflow-Builder orchestriert komplexe Aufgaben von automatisierter Kundenbetreuung bis zu Datenbank-Abfragen. Die visuelle Gestaltung macht fortgeschrittene KI-Anwendungen ohne Programmierung zugänglich.

4 Bereitstellungsoptionen

Vier Plattformen stehen zur Verfügung: **AWS** mit Bedrock-Integration, **Azure** mit OpenAI-Anbindung, **Google Cloud** mit Vertex AI, oder vollständig **On-Premise**. Die Standard-Installation nutzt Cloud-APIs der gewählten Plattform. Die Professional-Installation ermöglicht lokale Open-Source-Modelle für alle Plattformen einschließlich vollständig Air-gapped On-Premise-Deployments.

5 Zentrale Vorteile

Flexible Datenkontrolle: Wählen Sie zwischen vertrauenswürdigen Cloud-APIs (Standard) oder vollständiger lokaler Verarbeitung ohne externe Abhängigkeiten (Professional). Diese Flexibilität ermöglicht Anpassung an unterschiedliche Compliance-Anforderungen.

Nachvollziehbare KI: Jede Antwort verweist auf Quelldokumente. Die Visualisierung von Text-Chunks schafft Transparenz und Vertrauen in KI-generierte Informationen.

Skalierbare Implementation: Die Lösung wächst von Pilotgruppen zu unternehmensweiten Deployments. Die containerisierte Architektur unterstützt horizontale Skalierung ohne fundamentale Redesigns.

Workflow-Automatisierung: Der Agent-Builder ermöglicht komplexe Geschäftsprozesse ohne Programmierung - von Kundenbetreuung bis zu mehrstufigen Recherche-Workflows.

Anbieter-Unabhängigkeit: Die Standard-Version erlaubt Wechsel zwischen verschiedenen Cloud-Anbietern. Die Professional-Version bietet vollständige Unabhängigkeit durch eigene Open-Source-Modelle.

6 Überlegungen zur Bereitstellung

Für Standard-Installationen benötigen Sie administrativen Zugriff zur Cloud-Ressourcen-Provisionierung und API-Credentials für Ihren gewählten KI-Anbieter. Die Terraform-Automatisierung standardisiert die Infrastruktur-Einrichtung.

Professional-Installationen erfordern zusätzlich Infrastruktur für das Hosting eigener Modelle. Der Aufwand ist deutlich höher, ermöglicht aber vollständige Datenkontrolle. Die Identifikation relevanter Dokumentenquellen und Definition von Zugriffsberechtigungen sollten vor der technischen Implementierung geklärt sein.

Der Pilot validiert Technologie und etabliert Nutzungsmuster. Innerhalb von 2-4 Wochen kristallisieren sich Erkenntnisse über Performance-Anforderungen heraus, die eine informierte Skalierungsentscheidung ermöglichen.

7 Starten Sie jetzt

Technische Evaluation: Nutzen Sie die Dokumentation im GitHub Repository für eigenständige Evaluation.

Standard-Installation (699,-): Kontaktieren Sie mich für ein Deployment mit Cloud-API-Integration. Installation erfolgt nach Validierung an einem einzelnen Tag.

Professional-Lösung (1200,- pro Tag): Vereinbaren Sie einen Beratungstermin zur Diskussion Ihrer Anforderungen für vollständig lokale Modelle oder andere spezifische Integrationen.

Kontakt: info@anwi.gmbh — https://anwi.gmbh — https://github.com/andreaswittmann

8 Ressourcen

- RagFlow Projekt: https://github.com/infiniflow/ragflow
- Private RAG Repository: https://github.com/andreaswittmann/private-rag
- Private AI Chat-Lösung: https://anwi.gmbh/?page_id=1156&lang=de

Die Private AI Agent Solution ergänzt sich mit der Private AI Chat-Lösung für ein umfassendes KI-Ökosystem in Ihrer Infrastruktur.