Private AI Agent Solution: Knowledge Base, Chat and Workflow Automation

Andreas Wittmann

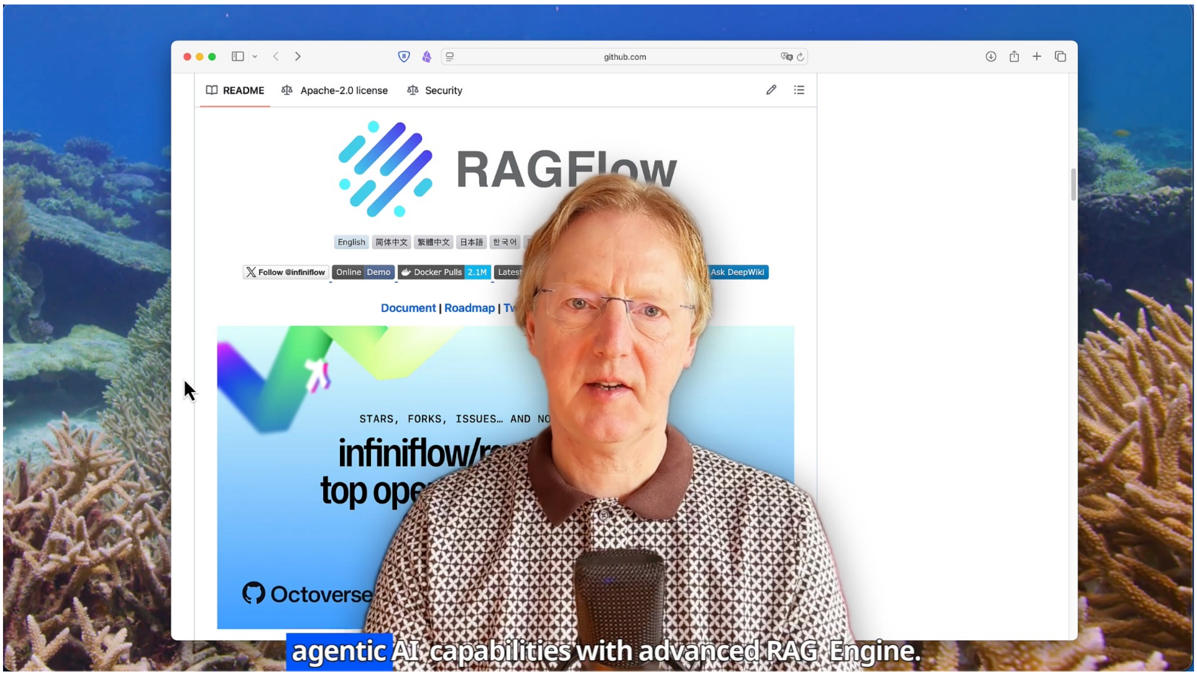
andreas.wittmann@mailbox.org

Datum: 2025-11-11

The Private AI Agent Solution is a fully configured RagFlow installation that combines a knowledge base, chat client and AI agents for the intelligent processing of your company documents – either with trusted cloud APIs (Standard, 699) or completely air-gapped with your own open source models (Professional, 1,200/day).

Download: PDF version of this document.

1 Introduction



This video shows the deployment of RAGFlow on AWS using the live scripting approach. The deployment utilises the open source project **private-rag**, which forms the basis for the **Starter** and **Standard** offerings. `https://www.youtube.com/watch?v=fLg-f9MkRMo`

The *Private AI Agent solution* is a fully configured installation of the open-source framework *RagFlow*. This solution combines a knowledge base, chat client and AI agents into an integrated platform for the intelligent processing of your company documents.

RagFlow processes various document formats through intelligent parsing and semantic indexing. The result is accurate answers with direct source references that ensure traceable and hallucination-free communication. The solution enables your teams to systematically tap into corporate knowledge – from interactive question-and-answer dialogues to automated workflows.

Designed for pilot groups of 5-30 users, the installation utilises my proven private-rag project. Four deployment options are available: AWS, Azure, Google Cloud or on-premise.

2 Three offer types

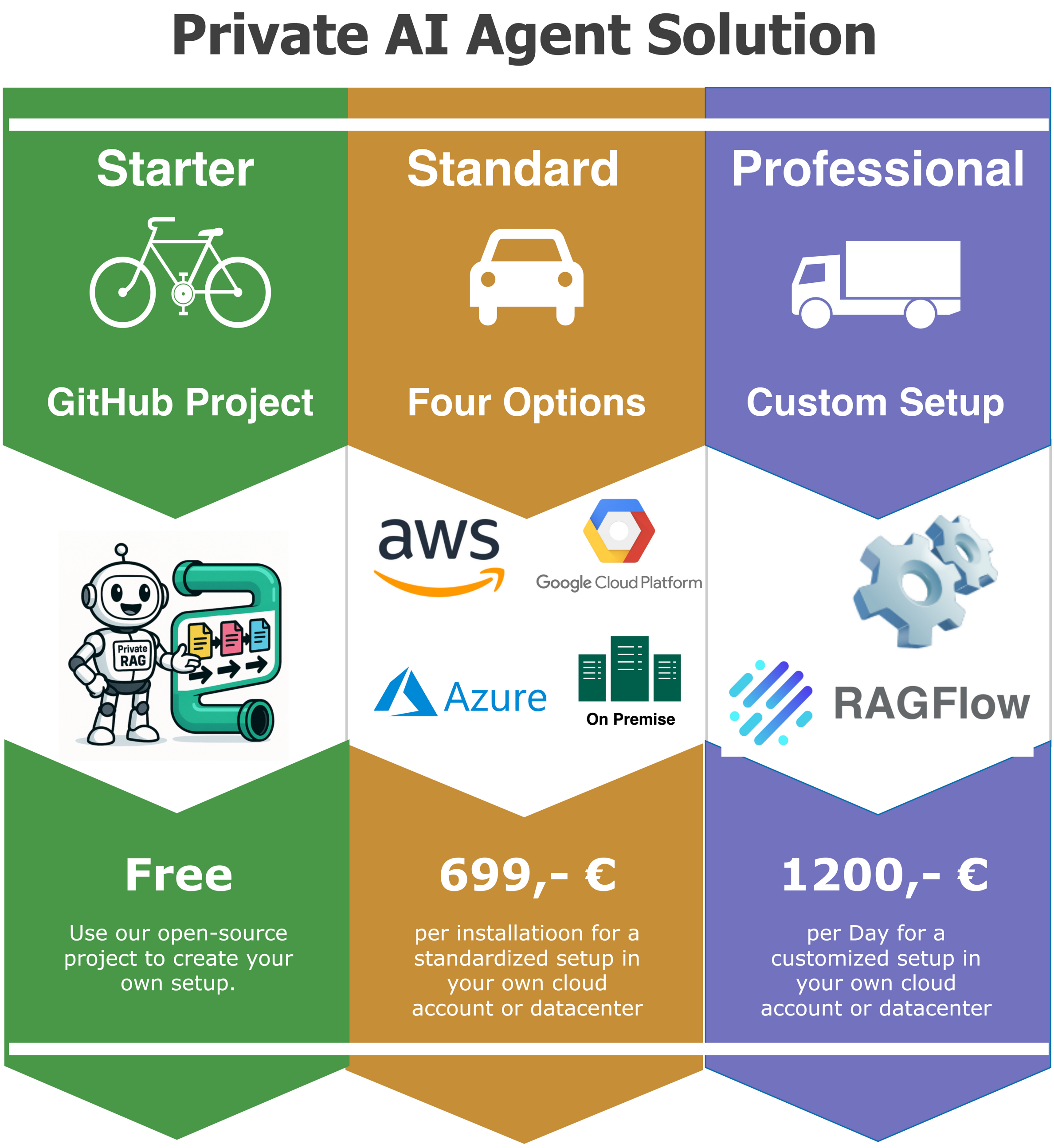

Private AI Agent Solution
Starter
GitHub Project
Standard
Four Options
Professional
Custom Setup
Private RAG
aws
Google Cloud Platform
Azure
On Premise
RAGFlow
Free
Use our open-source project to create your own setup.
699,- €
per installatioon for a standardized setup in your own cloud account or datacenter
1200,- €
per Day for a customized setup in your own cloud account or datacenter

Figure 1: Overview of Private AI Agent Solution offerings

2.1 Starter (free)

Technically savvy teams can access the complete deployment documentation via my GitHub repository. The step-by-step guide enables independent technology evaluation prior to commercial implementation.

2.2 Standard (699 per installation)

The Standard package provides a fully configured RagFlow installation with a knowledge base, chat client, and agent builder. Installation takes place in one day after a preparation phase and includes a Docker-containerised instance, SSL/HTTPS-secured access, Terraform-based setup, and complete documentation.

This configuration uses external AI APIs for language models and embeddings. Information is forwarded to the AI providers you select. You choose trusted providers such as AWS Bedrock, Azure OpenAI or Google Vertex AI based on your existing cloud infrastructure.

2.3 Professional (1,200 per day)

The Professional package allows you to use your own open-source models for complete data control. This installation can be operated as an air-gapped system without external API dependencies – all components operate exclusively within your infrastructure.

The effort required for this configuration is significantly higher, as your own embedding models and language models must be provided. Typical additional scenarios include integration into authentication systems, customisation of the document pipeline, performance optimisation, or migration of existing knowledge databases.

3 Core functions

Knowledge base: Intelligent parsing processes PDFs, Word, Excel, PowerPoint and other formats. Template-based chunking preserves semantic relationships. You control the processing by visualising the text chunks and can intervene manually if necessary.

Chat client: Natural language questions are answered with precise source references. Specialised assistants can be configured for different subject areas, each with its own knowledge base.

AI agents: The no-code workflow builder orchestrates complex tasks from automated customer support to database queries. The visual design makes advanced AI applications accessible without programming.

4 Deployment options

Four platforms are available: **AWS** with Bedrock integration, **Azure** with OpenAI connection, **Google Cloud** with Vertex AI, or completely **on-premise**. The standard installation uses cloud APIs from the selected platform. The professional installation enables local open-source models for all platforms, including fully air-gapped on-premise deployments.

5 Key advantages

Flexible data control: Choose between trusted cloud APIs (Standard) or complete local processing without external dependencies (Professional). This flexibility allows adaptation to different compliance requirements.

Transparent AI: Each response references source documents. The visualisation of text chunks creates transparency and trust in AI-generated information.

Scalable implementation: The solution grows from pilot groups to company-wide deployments. The containerised architecture supports horizontal scaling without fundamental redesigns.

Workflow automation: The agent builder enables complex business processes without programming – from customer support to multi-stage research workflows.

Vendor independence: The standard version allows switching between different cloud providers. The professional version offers complete independence through its own open-source models.

6 Deployment considerations

For standard installations, you need administrative access to cloud resource provisioning and API credentials for your chosen AI provider. Terraform automation standardises infrastructure setup.

Professional installations additionally require infrastructure for hosting your own models. The effort involved is significantly higher, but it allows for complete data control. The identification of relevant document sources and the definition of access permissions should be clarified before technical implementation.

The pilot validates technology and establishes usage patterns. Within 2-4 weeks, insights into performance requirements emerge, enabling an informed scaling decision.

7 Get started now

Technical evaluation: Use the documentation in the GitHub repository for independent evaluation.

Standard installation (699): Contact me for deployment with Cloud API integration. Installation takes place on a single day after validation.

Professional solution (1,200 per day): Schedule a consultation to discuss your requirements for fully local models or other specific integrations.

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

8 Resources

- RagFlow project: `https://github.com/infiniflow/ragflow`
- Private RAG repository: `https://github.com/andreaswittmann/private-rag`
- Private AI chat solution: `https://anwi.gmbh/?page_id=1107&lang=en`

The Private AI Agent Solution complements the Private AI Chat Solution for a comprehensive AI ecosystem in your infrastructure.