CHATBOT LEI DO BEM , MATECH SOLUTIONS

📄 Manual de Implementação Detalhado: Chatbot RAG (n8n)

Guia de Implantação e Configuração

Solução: Chatbot com Geração Aumentada por Recuperação (RAG)

Plataforma: n8n - Orquestração e Automação

---

1. Introdução e Arquitetura RAG

Este documento é o seu guia para implementar um **Chatbot RAG (Generation Augmented Retrieval)** eficiente. O sistema utiliza o **n8n** para orquestrar a busca em sua base de conhecimento (documentos internos) antes de gerar a resposta final através de um LLM.

O Fluxo RAG em Ação

1. **Entrada:** A pergunta do usuário é recebida pelo **Webhook** do n8n.
2. **Busca (R):** A consulta é vetorizada e enviada ao Banco de Dados Vetorial para **recuperar** o contexto relevante.
3. **Geração (AG):** O n8n envia o contexto recuperado junto com a pergunta original ao LLM, instruindo-o a gerar uma resposta **factual e fundamentada**.

---

2. Pré-requisitos e Credenciais

Antes de começar, confirme que possui acesso a todos os serviços externos necessários.

Serviço Necessário	Função no Sistema RAG	Credenciais Essenciais
Instância n8n	Motor de automação para orquestrar o fluxo.	Acesso de Login/Desenvolvedor.
Serviço de LLM	Geração de texto e respostas.	**Chave de API** (API Key).
Serviço de Embeddings	Conversão de texto em representações vetoriais.	**Chave de API** (Pode ser a mesma do LLM).
Banco de Dados Vetorial	Armazenamento e busca semântica de documentos.	Host URL, Chave de API, Nome do Índice.
Endpoint Público	Necessário para expor o Webhook do n8n à internet.	Depende do modelo de implantação (item 3).

---

3. ☁️ Modelos de Implantação do n8n

O primeiro passo para o funcionamento é decidir onde seu n8n será executado e como o *Webhook* (o ponto de entrada das perguntas) será exposto publicamente.

Opção A: n8n Cloud (Recomendado para Facilidade)

Se você utiliza o n8n Cloud, a infraestrutura e a exposição pública do *Webhook* são gerenciadas automaticamente:

1. **Login:** Acesse sua conta n8n Cloud.
2. **URL Estável:** O n8n Cloud garante um URL público e estável para todos os *webhooks*. Nenhuma configuração adicional de rede ou túnel é necessária.
3. **Continuação:** Prossiga para o **Capítulo 4**.

Opção B: Máquina Local ou Self-Hosted (Requer Configuração de Rede)

Se você roda o n8n em sua máquina local ou em um servidor privado sem acesso público direto, você deve criar um túnel para expor o *Webhook* à internet.

Para o desenvolvimento desse back-end, foi utilizada essa opção, que facilita a depuração dos testes.

Passo 1: Instalação do Túnel

Ferramentas como o **ngrok** criam um túnel seguro de um endereço local (ex: `http://localhost:5678`) para um endereço público temporário na internet.

1. **Instale o ngrok:** Baixe e instale o ngrok (ou ferramenta similar).
2. **Autentique:** Configure seu token de autenticação (disponível após o registro no ngrok).
3. **Execute o Túnel:** Abra seu terminal e execute o comando, substituindo a porta se necessário:
   ```
   ngrok http 5678
   ```
4. **Obtenha a URL:** O ngrok fornecerá um URL público (ex: `https://abcd1234efgh.ngrok-free.app`). **Anote este endereço.**

Passo 2: Configuração da URL Base no n8n

Você deve informar ao n8n qual é a URL pública que ele deve usar para construir o *endpoint* do *Webhook*:

1. **Variável de Ambiente:** Configure a variável de ambiente `WEBHOOK_URL` da sua instância n8n com o endereço público fornecido pelo ngrok:
   ```
   WEBHOOK_URL=[https://abcd1234efgh.ngrok-free.app]
   (https://abcd1234efgh.ngrok-free.app)
   ```
2. **Reinicie:** Reinicie sua instância do n8n para que a nova `WEBHOOK_URL` seja aplicada.

---

4. 📝 Importação e Configuração do Workflow

Com o n8n em execução (e com a `WEBHOOK_URL` configurada, se for o caso), o próximo passo é importar e configurar o fluxo.

4.1. Importação do Arquivo

1. **Localize o JSON:** O arquivo `chatbot_rag_workflow.json` (ou nome fornecido) contém toda a lógica do chatbot.
2. **No n8n:** Crie um novo *workflow* e utilize a opção **Import from JSON**.
3. **Confirmação:** O *workflow* completo será carregado e estará **inativo**.

4.2. Vinculação de Credenciais e APIs

O *workflow* importado deve ser conectado aos seus serviços.

1. **Edite os Nós de Serviço:** Clique duas vezes em cada nó que representa um serviço externo (`LLM Node` , `DB Vector Node` , etc.).
2. **Criação/Seleção:** No painel lateral, no campo **"Credential"** (Credencial):
   - **Se for a primeira vez:** Clique em **"Create New"** (Criar Nova) e insira suas chaves de API/tokens de acesso.
   - **Se já existir:** Selecione a credencial criada previamente.
3. **Repetição:** Repita o processo até que todos os nós externos estejam vinculados às suas credenciais de produção.

---

5. 🎯 Ajustes Finos (Otimização RAG)

Revise e ajuste os parâmetros a seguir para otimizar o desempenho do RAG para o seu conjunto de documentos.

Ajustes de Busca (Nó DB Vetorial)

Parâmetro	Descrição Detalhada	Sugestão de Configuração
Index Name	Nome exato da coleção/índice no DB Vetorial onde seus documentos estão armazenados. **Deve ser idêntico ao nome usado na indexação.**	*Ex:* `documentos-tecnicos-matech`
Top K	Quantidade de documentos mais relevantes que o LLM usará como contexto para responder. Valores mais altos aumentam a precisão, mas podem custar mais e aumentar a latência.	3 a 5

Ajustes de Geração (Nó LLM)

Parâmetro	Descrição Detalhada	Importância
Model	Nome do modelo LLM escolhido (GPT-4o, Gemini 2.5 Flash, etc.). A escolha afeta o custo e a qualidade da resposta.	Alta
System Prompt	Instrução de alto nível que define o papel, o tom e as restrições do chatbot. **É crucial instruir o LLM a usar apenas o contexto fornecido.**	Crítica
Temperature	Controla a aleatoriedade da resposta (0 = factual e previsível; 1 = criativo e variado). Mantenha baixo para RAG.	Baixa (0.1 a 0.5)

6. Ativação e Teste Final

Com todas as conexões e configurações aplicadas, o *workflow* está pronto para ser ativado.

6.1. Ativação do Workflow

1. **Localize o Toggle:** No topo da tela do n8n, encontre o interruptor de status.
2. **Ativar:** Mude o interruptor para o estado **"Active"** (Ativo).

6.2. Obtenção do Endpoint de Produção

1. **Nó Webhook:** Clique no nó inicial do *workflow* (`Webhook`).
2. **Copie o URL:** O n8n gerará um URL. Este é o **Endpoint de Produção** oficial do seu chatbot.
   - **Cloud:** Será um URL estável do n8n.
   - **Local/ngrok:** Será a URL pública do ngrok que você configurou no **Capítulo 3**.

6.3. Teste de Validação

Para garantir o sucesso da implementação, envie uma pergunta:

1. Use uma ferramenta de teste (Postman) ou seu próprio aplicativo para enviar uma requisição `POST` ao URL de produção do *Webhook*.
2. **Valide o Fluxo:** Verifique no n8n se o *workflow* foi executado.
3. **Valide a Resposta:** Confirme que a resposta final é **fiel** e **fundamentada** no seu conjunto de documentos.

Sucesso! A implementação do seu Chatbot RAG no n8n está completa. Lembre-se, alguns nós contidos no webhook exigem credenciais! Por ser informações privadas, o ideal é que você mesmo "logue" com as próprias credenciais.