

OGR

+ RAG Híbrido



Mistral AI



supabase



VILA AMETISTA - TORRE 01 - APTO 210 VI

The floor plan displays the layout of the apartment, including rooms labeled 'SALA', 'COZINHA', 'QUARTO', 'QUARTO', and 'BANHEIRO'. A legend on the right side lists the tracked objects:

- Porta aberta em cozinha
- Caixa de lixo que caiu no chão
- Lata
- Porta aberta em banheiro infantil
- Porta aberta em cozinha
- Anel de rosca na lata
- Lata no chão da cozinha
- Lata
- Porta aberta em cozinha



Tecnologia OCR: Uma Visão Abrangente

A tecnologia de Reconhecimento Ótico de Caracteres (OCR) é um campo inovador que permite a conversão de diferentes tipos de documentos, como papéis digitalizados, PDFs ou imagens, em dados editáveis e pesquisáveis. Essa tecnologia vai muito além da simples digitalização, transformando imagens estáticas de texto em formatos que computadores podem entender e manipular.

Como Funciona?

O processo de OCR envolve várias etapas sofisticadas. Primeiramente, a imagem do documento é analisada para identificar a estrutura da página, como blocos de texto, imagens e tabelas. Em seguida, o software OCR detecta os caracteres individuais, usando algoritmos para comparar os padrões encontrados com fontes conhecidas ou para identificar formas de caracteres. Uma vez reconhecidos, esses caracteres são convertidos em texto digital que pode ser editado, copiado e pesquisado.

Aplicações Práticas

- **Automação de Escritório:** Digitalização e indexação automática de documentos, faturas, contratos e recibos.
- **Saúde:** Conversão de registros médicos em papel para prontuários eletrônicos, facilitando o acesso e a gestão de dados de pacientes.
- **Setor Jurídico:** Transformação de documentos legais em formato digital pesquisável, agilizando processos de pesquisa e auditoria.
- **Educação:** Criação de materiais de estudo acessíveis para alunos com deficiência visual, convertendo textos impressos em formatos digitais.
- **Bancos e Finanças:** Processamento de cheques, formulários e outros documentos financeiros para otimizar operações e reduzir erros.

Benefícios do OCR

- **Eficiência Aumentada:** Reduz significativamente o tempo gasto com a entrada manual de dados.
- **Redução de Custos:** Diminui a necessidade de mão de obra para tarefas repetitivas e o uso de papel.
- **Precisão Otimizada:** Minimiza erros humanos, resultando em dados mais confiáveis.
- **Pesquisa Facilitada:** Torna documentos arquivados facilmente pesquisáveis e acessíveis.
- **Melhor Gestão Documental:** Contribui para um ambiente de trabalho mais organizado e com melhor fluxo de informações.
- **Sustentabilidade:** Apoia a transição para ambientes de trabalho sem papel, contribuindo para a redução do impacto ambiental.



Mistral AI: Inovação em Modelos de Linguagem

A Mistral AI é uma empresa francesa de inteligência artificial que rapidamente se estabeleceu como uma força inovadora no campo dos Large Language Models (LLMs). Fundada por ex-pesquisadores do Google DeepMind e Meta, a Mistral dedica-se a desenvolver modelos de IA de código aberto, eficientes e de alto desempenho, que oferecem alternativas robustas aos modelos proprietários existentes.

Características Principais

- Modelos Abertos e Transparentes:** Um dos pilares da Mistral AI é o compromisso com a natureza open-source de seus modelos. Isso promove a transparência, a colaboração da comunidade e permite que desenvolvedores e empresas personalizem e otimizem os modelos para suas necessidades específicas.
- Eficiência e Desempenho:** A Mistral se destaca por criar LLMs que, embora menores em tamanho, demonstram desempenho comparável ou superior a modelos muito maiores em diversas tarefas. Isso resulta em inferências mais rápidas e custos computacionais reduzidos.
- Inovação Arquitetônica:** A empresa tem sido pioneira em arquiteturas como o Sparse Mixture-of-Experts (SMoE) no seu modelo Mixtral, que permite uma utilização mais eficiente dos recursos computacionais.
- Flexibilidade e Escalabilidade:** Os modelos da Mistral são projetados para serem altamente adaptáveis, podendo ser implantados em diferentes ambientes, desde dispositivos locais até infraestruturas de nuvem complexas.

Modelos Disponíveis

- Mistral 7B:** Um dos primeiros e mais populares modelos da empresa, o Mistral 7B é um LLM de 7 bilhões de parâmetros que oferece excelente desempenho para uma vasta gama de aplicações, sendo ideal para tarefas que exigem eficiência e baixa latência.
- Mixtral 8x7B:** Este modelo utiliza a arquitetura Mixture-of-Experts, onde 8 "especialistas" de 7 bilhões de parâmetros trabalham em conjunto. Isso permite que o modelo tenha um desempenho comparável a modelos muito maiores (como um de 129B de parâmetros) com uma fração dos custos computacionais durante a inferência.
- Mistral Large:** Lançado para competir diretamente com os modelos mais avançados do mercado, o Mistral Large representa um passo significativo em termos de capacidade e complexidade, oferecendo desempenho de ponta para as aplicações mais exigentes.

Aplicações Práticas

- Geração de Texto Criativo:** Criação de artigos, roteiros, e-mails e outros conteúdos escritos.
- Assistentes Virtuais e Chatbots:** Desenvolvimento de sistemas de conversação mais naturais e eficientes.
- Sumarização de Documentos:** Extração de pontos-chave de textos longos para economizar tempo.
- Tradução Automática:** Serviços de tradução de alta qualidade para múltiplos idiomas.
- Geração e Análise de Código:** Auxílio a desenvolvedores na escrita, depuração e otimização de código.
- Análise de Dados e Insights:** Processamento de grandes volumes de dados para identificar tendências e padrões.

Benefícios da Plataforma Mistral

- Otimização de Custos:** Modelos mais eficientes significam menor consumo de recursos computacionais, resultando em economia para empresas e desenvolvedores.
- Maior Controle e Personalização:** A natureza open-source permite que os usuários adaptem os modelos às suas necessidades específicas, garantindo maior controle sobre a IA.
- Desempenho de Ponta:** A Mistral oferece modelos que competem com os líderes do setor em termos de qualidade e capacidade.
- Comunidade Ativa e Suporte:** O ecossistema open-source fomenta uma comunidade de desenvolvedores ativa, contribuindo para a evolução contínua dos modelos e oferecendo suporte mútuo.
- Segurança e Privacidade:** Para muitas empresas, a capacidade de rodar modelos on-premise ou com maior controle sobre os dados é um benefício crucial para a segurança e conformidade.

O RAG Híbrido: Uma Evolução na Geração Aumentada por Recuperação

O RAG híbrido (Retrieval-Augmented Generation) representa um avanço significativo sobre o RAG tradicional, combinando múltiplas estratégias de recuperação de informações para fornecer respostas mais precisas, abrangentes e contextualmente relevantes. Ele integra as forças de diferentes métodos de busca para otimizar a forma como os Modelos de Linguagem Grandes (LLMs) acessam e utilizam dados externos.

Como Funciona o RAG Híbrido?

Diferente do RAG tradicional, que geralmente se baseia apenas na recuperação densa (como embeddings vetoriais para similaridade semântica), o RAG híbrido incorpora uma combinação de abordagens de recuperação. As mais comuns incluem:

- **Recuperação Densa (Vector Search):** Utiliza embeddings para encontrar documentos semanticamente semelhantes à consulta, ideal para capturar o significado contextual.
- **Recuperação Esparsa (Keyword Search):** Emprega métodos baseados em palavras-chave (como BM25 ou TF-IDF) para encontrar correspondências lexicais diretas, sendo eficaz para termos específicos ou nomes próprios.

Ao combinar essas estratégias, o sistema RAG híbrido pode primeiro realizar várias buscas em paralelo ou sequencialmente. Os resultados de cada método são então agregados e frequentemente passam por uma etapa de re-ranking, onde um algoritmo mais sofisticado avalia a relevância combinada dos documentos recuperados, garantindo que o contexto mais rico e pertinente seja entregue ao LLM para a geração da resposta.

Diferenças do RAG Tradicional

Enquanto o RAG tradicional aprimora os LLMs com informações externas, sua dependência exclusiva da recuperação densa pode ter limitações. O RAG híbrido supera essas limitações ao:

- **Ampla Gama de Consultas:** Lidar de forma mais eficaz com consultas que exigem tanto compreensão semântica profunda quanto precisão lexical (ex: "Qual o CNPJ da empresa X?" vs. "Quais são as melhores práticas de sustentabilidade?").
- **Robustez a Dados Ruim:** Ser mais resiliente a dados de baixa qualidade ou consultas mal formuladas, pois a combinação de métodos aumenta a chance de recuperar informações relevantes.
- **Superar Desafios de Embeddings:** Mitigar problemas como "out-of-vocabulary" (OOM) ou a incapacidade de embeddings de capturar nuances específicas de alguns termos.

Vantagens do RAG Híbrido

- **Melhora na Qualidade da Resposta:** Produz saídas mais precisas e informativas ao acessar um espectro mais amplo de dados relevantes.
- **Redução de Alucinações:** Diminui significativamente a tendência dos LLMs de gerar informações incorretas ou inventadas.
- **Maior Cobertura de Informações:** Garante que tanto documentos com correspondência exata de termos quanto aqueles com similaridade conceitual sejam considerados.
- **Flexibilidade e Adaptabilidade:** Pode ser ajustado para diferentes tipos de dados e domínios, desde documentos técnicos até bases de conhecimento gerais.

Aplicações Práticas

O RAG híbrido é ideal para cenários onde a precisão e a completude das informações são críticas:

- **Atendimento ao Cliente e Chatbots:** Fornecer respostas rápidas e precisas a uma vasta gama de perguntas dos usuários.
- **Pesquisa Jurídica e Médica:** Ajudar profissionais a encontrar precedentes legais ou informações clínicas com alta especificidade e relevância.
- **Bases de Conhecimento Empresariais:** Permitir que funcionários acessem informações internas complexas de forma eficiente.
- **Sistemas de Perguntas e Respostas Complexos:** Abordar questões que exigem tanto uma compreensão profunda do tema quanto a recuperação de fatos específicos.

Benefícios da Abordagem Híbrida

A adoção do RAG híbrido traz uma série de benefícios estratégicos:

- **Eficiência Aumentada:** Otimiza o tempo de recuperação e a relevância das informações, tornando os sistemas mais eficientes.
- **Melhor Experiência do Usuário:** Oferece respostas mais satisfatórias e confiáveis, aumentando a confiança no sistema.
- **Escalabilidade:** Permite lidar com grandes volumes de dados de forma mais robusta e eficaz.
- **Vantagem Competitiva:** Empresas que implementam RAG híbrido podem oferecer soluções de IA conversacional e de busca de conhecimento superiores.

Apresentação

Conheça um pouco mais sobre a solução

Criando uma chave de API da Mistral

Todos os passos para criar sua chave de api.

Configurando o fluxo

Importando os Fluxos para o n8n e configurando as credenciais.

1

2

3

4

5

6

Requerimentos

Tudo que você vai precisar para começar

Configurando o Supabase

Configure sua conta no Supabase

Suporte

Entenda como pedir suporte

Extração de Dados de PDFs com Mistral OCR para Supabase RAG

Automatize a extração inteligente de **texto e imagens** de arquivos PDF e envie os dados diretamente para um **banco de dados Supabase**, pronto para ser usado em sistemas baseados em **RAG (Retrieval-Augmented Generation)**.

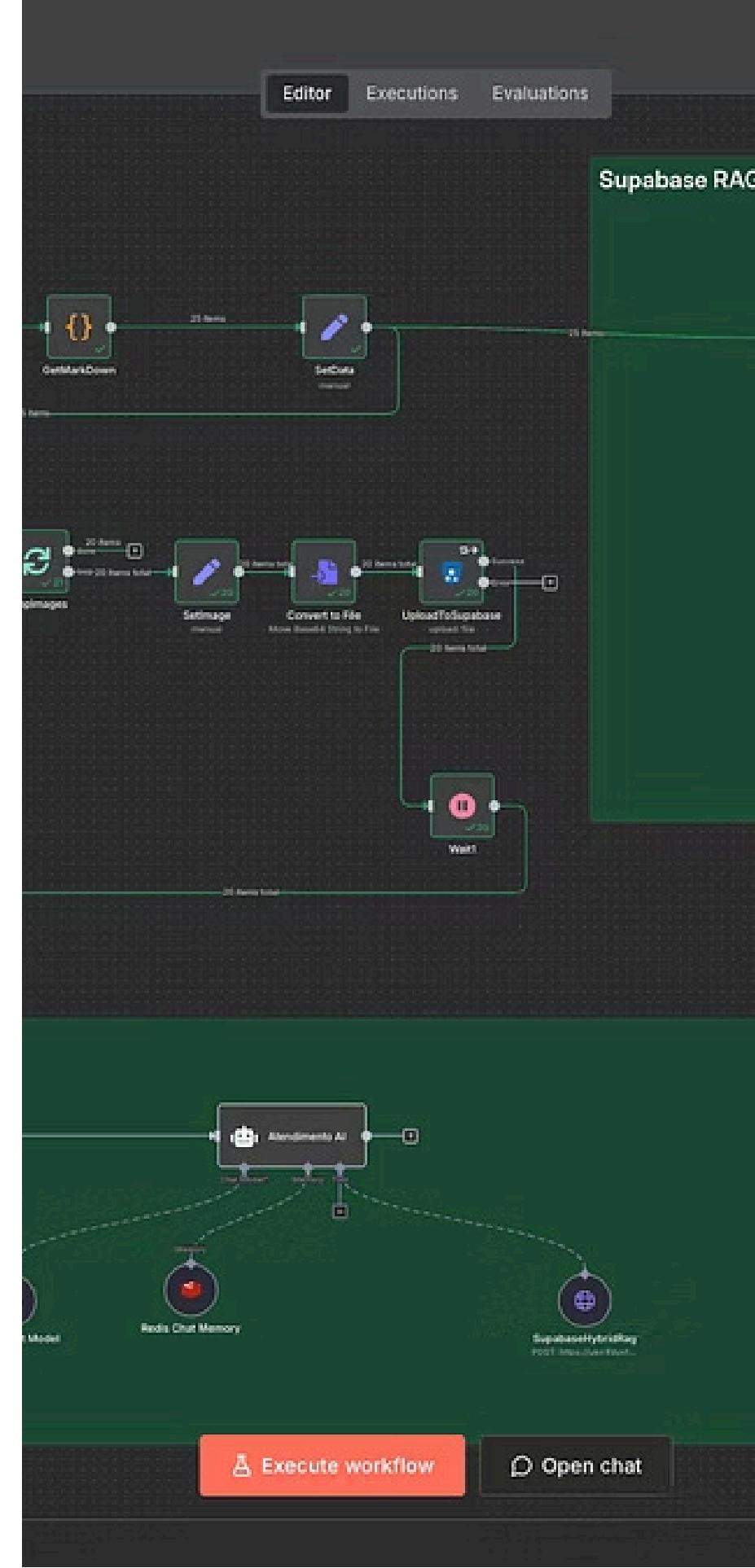
Este fluxo foi desenvolvido para lidar com documentos complexos — como contratos, relatórios e PDFs com tabelas ou imagens — usando o poder do **Mistral OCR** e a flexibilidade do **n8n**.

✓ O que este fluxo faz:

- 🧠 Usa **Mistral OCR** para extrair texto e imagens de arquivos PDF
- 🔄 Organiza e transforma os dados automaticamente
- 💾 Insere o conteúdo estruturado em uma tabela no **Supabase**
- 🔗 Pronto para integração com pipelines de **RAG híbrido**

🚀 Benefícios:

- Elimine o trabalho manual de leitura e digitação de PDFs
- Acelere ingestão de dados para uso em IA generativa
- Reduza erros humanos em processos de automação
- Totalmente customizável e documentado



Talk to the Chat



Get a demo >

Start building >

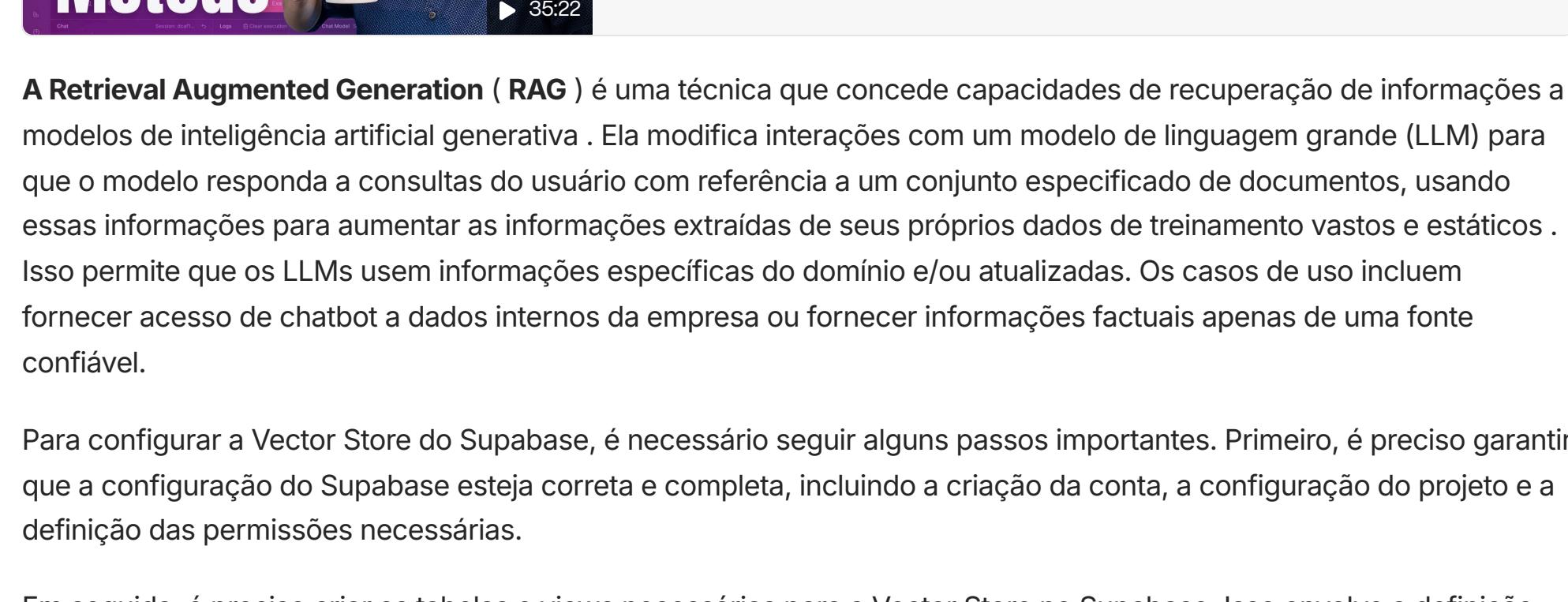
Criando uma Chave de API na Mistral

Acesse a plataforma mistral.ai e crie sua conta de testes. Preencha todos os dados da organização e-mail de contato e todas as informações solicitadas.

Chave de API

Siga os passos a seguir para criar sua chave de API. Depois guarde-a em local seguro pois após a visualização ela não será mais exibida.

Configurando o RAG Híbrido no Supabase



A Retrieval Augmented Generation (RAG) é uma técnica que concede capacidades de recuperação de informações a modelos de inteligência artificial gerativa. Ela modifica interações com um modelo de linguagem grande (LLM) para que o modelo responda a consultas do usuário com referência a um conjunto especificado de documentos, usando essas informações para aumentar as informações extraídas de seus próprios dados de treinamento vastos e estáticos. Isso permite que os LLMs usem informações específicas do domínio e/ou atualizadas. Os casos de uso incluem fornecer acesso de chatbot a dados internos da empresa ou fornecer informações factuais apenas de uma fonte confiável.

Para configurar a Vector Store do Supabase, é necessário seguir alguns passos importantes. Primeiro, é preciso garantir que a configuração do Supabase esteja correta e completa, incluindo a criação da conta, a configuração do projeto e a definição das permissões necessárias.

Em seguida, é preciso criar as tabelas e views necessárias para a Vector Store no Supabase. Isso envolve a definição das estruturas de dados que serão utilizadas para armazenar e recuperar os dados de forma eficiente.

Passos iniciais:

1. Crie uma conta no Supabase
2. Crie um organização
3. Crie um Projeto no Supabase

Credenciais:
Copie e cole a URL do Supabase e a credencial de acesso.

Siga as instruções contidas no site

Copie e cole o código abaixo para gerar as tabelas no Banco de dados Vetorial Supabase:

Este script deve ser executado no SQL EDITOR do Supabase para criar a tabela 'documents' no Supabase e algumas functions e views necessárias para o Vector Store funcionar.

```
create table documents (
    id bigint primary key generated always as identity,
    content text,
    metadata json,
    fts tsvector generated always as (to_tsvector('english', content)) stored,
    embedding vector(1536)
);
```

Em seguida, criaremos índices nas fts colunas embedding para que suas consultas individuais permaneçam rápidas em escala:

```
-- Create an index for the full-text search
create index on documents using gin(fts);

-- Create an index for the semantic vector search
create index on documents using hnsw (embedding vector_ip_ops);
```

Por fim, criaremos nossa hybrid_search função:

```
create or replace function hybrid_search(
    query_text text,
    query_embedding vector(1536),
    match_count int,
    full_text_weight float = 1,
    semantic_weight float = 1,
    rff_k int = 50
)
returns setof documents
language sql
as $$
with full_text as (
    select
        id,
        -- Note: ts_rank_cd is not indexable but will only rank matches of the where clause
        -- which shouldn't be too big
        row_number() over(order by ts_rank_cd(fts, websearch_to_tsquery(query_text)) desc) as rank_ix
    from
        documents
    where
        fts @@ websearch_to_tsquery(query_text)
    order by rank_ix
    limit least(match_count, 30) * 2
),
semantic as (
    select
        id,
        row_number() over (order by embedding <#> query_embedding) as rank_ix
    from
        documents
    order by rank_ix
    limit least(match_count, 30) * 2
)
select
    documents.*
from
    full_text
    full outer join semantic
        on full_text.id = semantic.id
join documents
    on coalesce(full_text.id, semantic.id) = documents.id
order by
    coalesce(1.0 / (rff_k + full_text.rank_ix), 0.0) * full_text_weight +
    coalesce(1.0 / (rff_k + semantic.rank_ix), 0.0) * semantic_weight
desc
limit
least(match_count, 30)
$$;
```

Agora vamos criar uma [Função Edge](#) do Supabase para receber as consultas do agente de IA.

```
import { createClient } from 'npm:@supabase/supabase-js@2'
import OpenAI from 'npm:openai'

const supabaseUrl = Deno.env.get('SUPABASE_URL')
const supabaseServiceRoleKey = Deno.env.get('SUPABASE_SERVICE_ROLE_KEY')
const openaiApiKey = Deno.env.get('OPENAI_API_KEY')

Deno.serve(async (req) => {
    // Grab the user's query from the JSON payload
    const { query } = await req.json()

    // Instantiate OpenAI client
    const openai = new OpenAI({ apiKey: openaiApiKey })

    // Generate a one-time embedding for the user's query
    const embeddingResponse = await openai.embeddings.create({
        model: 'text-embedding-3-large',
        input: query,
        dimensions: 1536,
    })

    const [{ embedding }] = embeddingResponse.data

    // Instantiate the Supabase client
    // (replace service role key with user's JWT if using Supabase auth and RLS)
    const supabase = createClient(supabaseUrl, supabaseServiceRoleKey)

    // Call hybrid_search Postgres function via RPC
    const { data: documents } = await supabase.rpc('hybrid_search', {
        query_text: query,
        query_embedding: embedding,
        match_count: 10,
    })

    return new Response(JSON.stringify(documents), {
        headers: { 'Content-Type': 'application/json' },
    })
})
```

Incluido a sua credencial da OpenAI na [Função Edge](#)

Função Edge

Depois de criar a função Edge é necessário cadastrar sua chave de acesso da OpenAI criando uma variável 'OPENAI_API_KEY'.

Storage

Na guia Storage crie sua credencial de acesso para fazer um upload das imagens

Storage

O Supabase tem um storage padrão S3 que usaremos armazenar as imagens extraídas do Mistral OCR.

S3 Connection

Connect to your bucket using any S3-compatible service via the S3 protocol

Enable connection via S3 protocol

Endpoint: https://s3.us-east-1.amazonaws.com/your-bucket-name

Region: sa-east-1

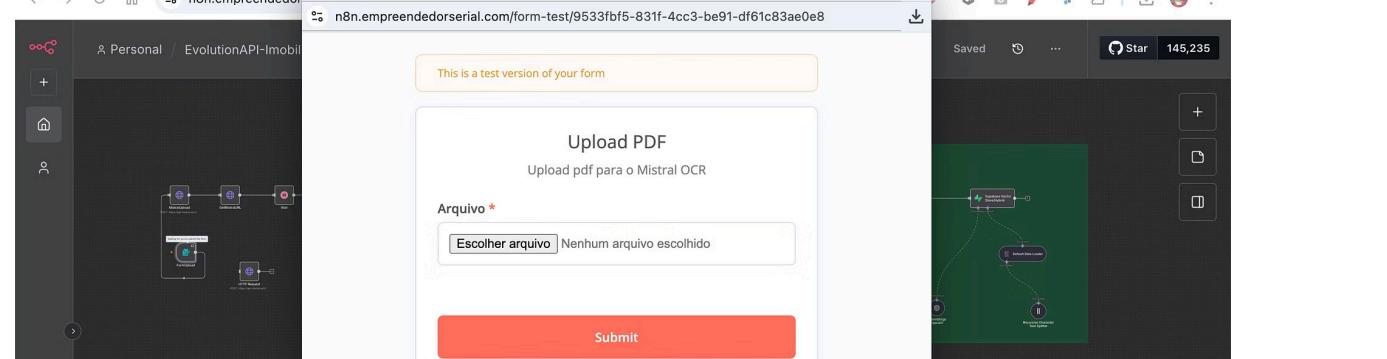
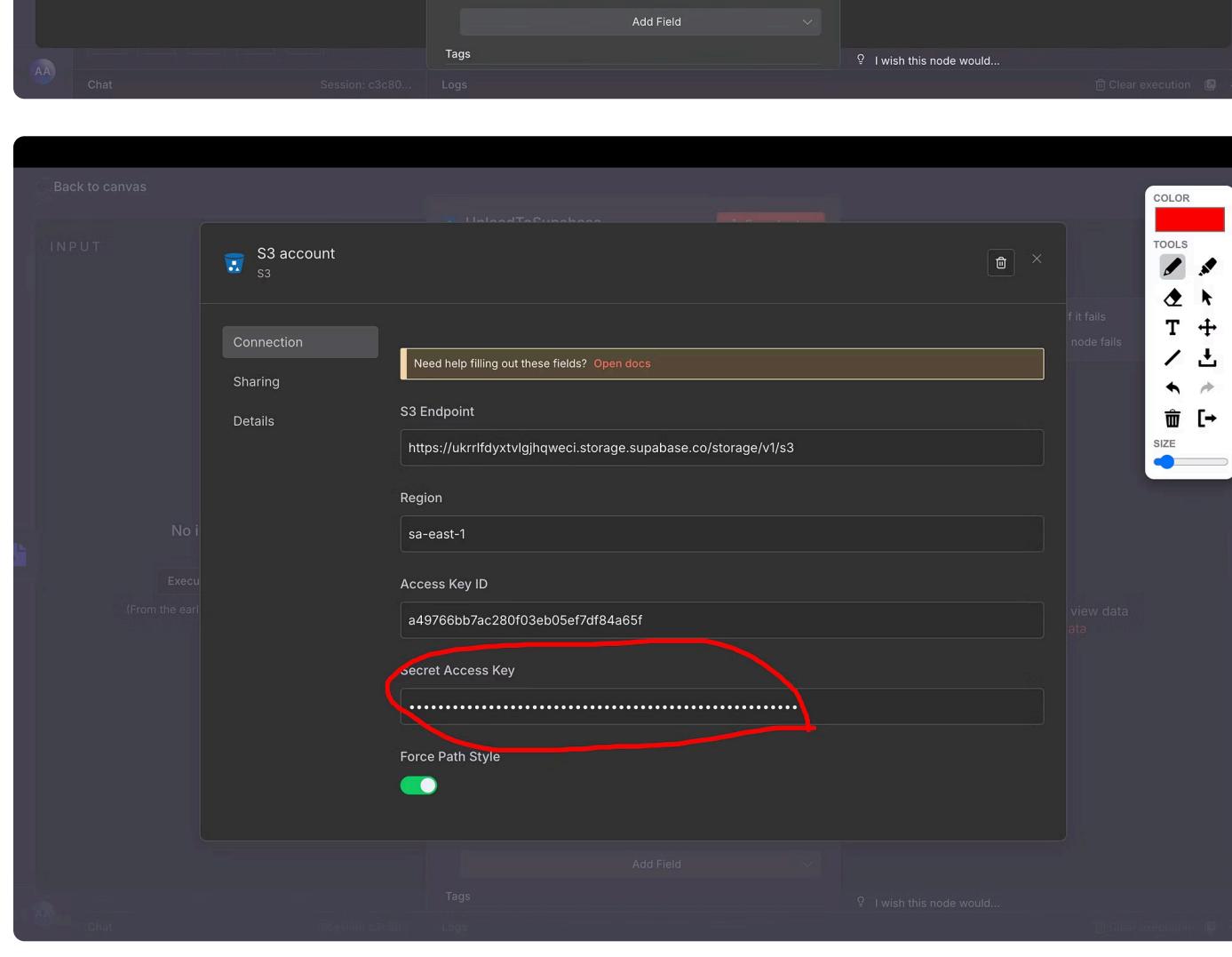
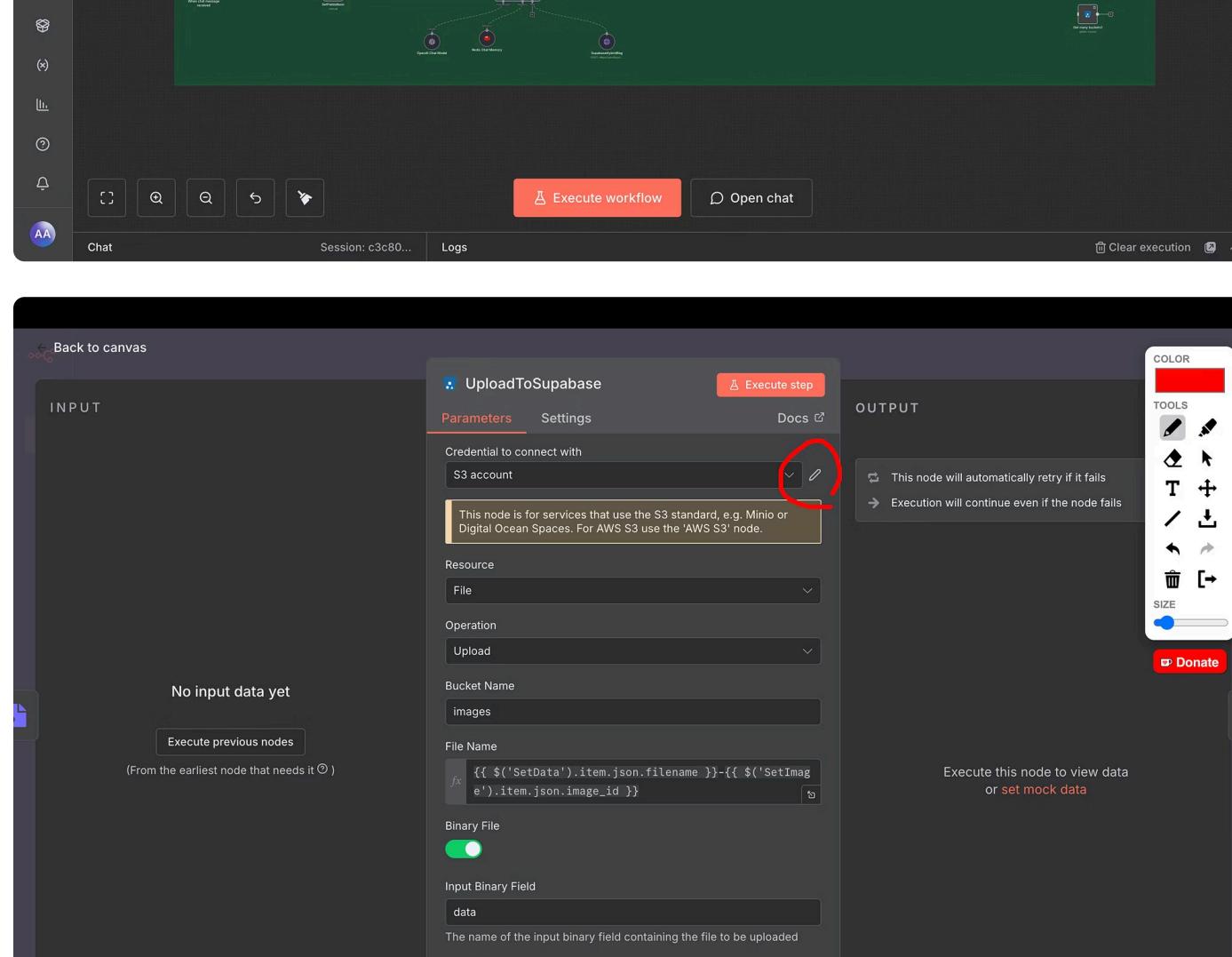
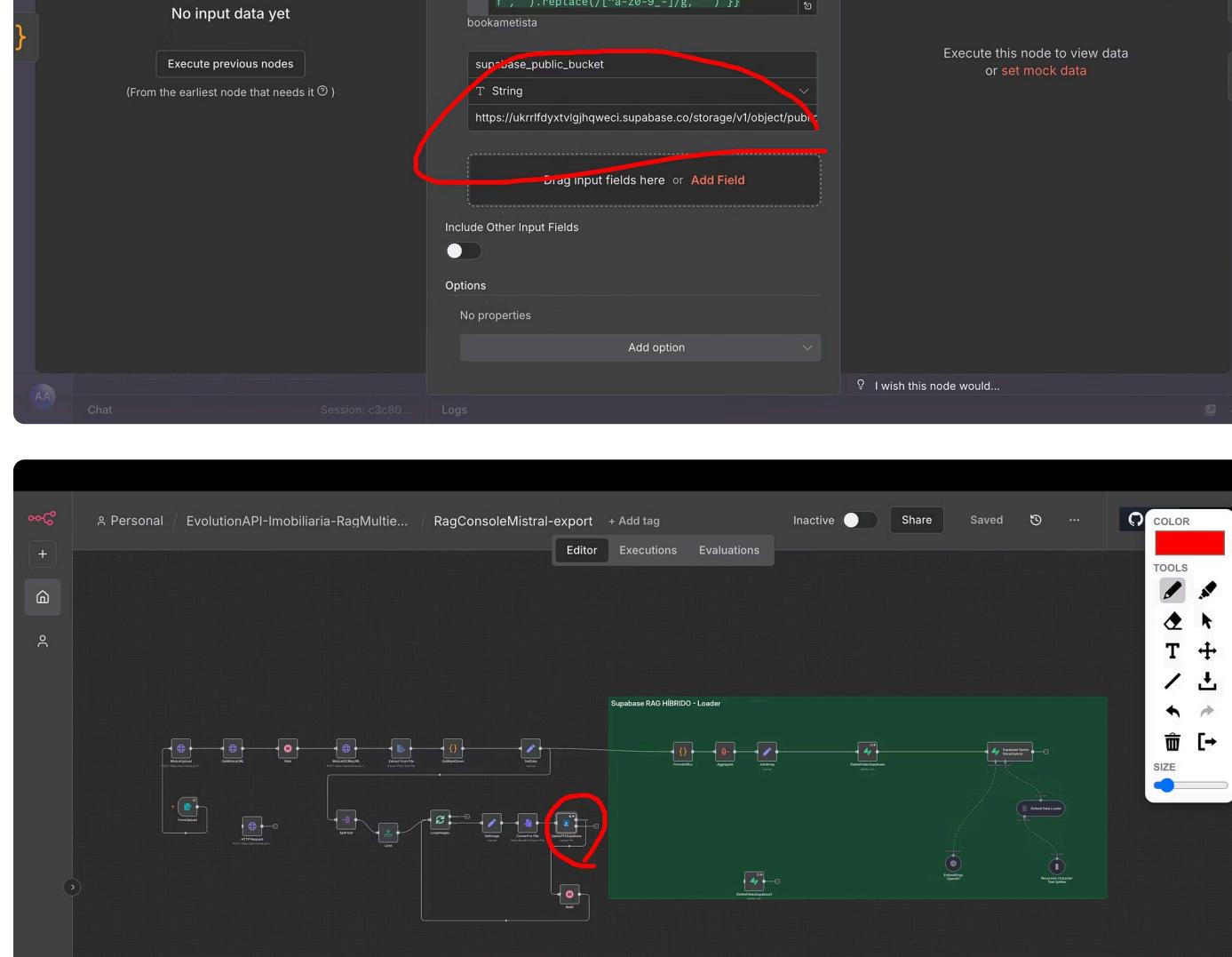
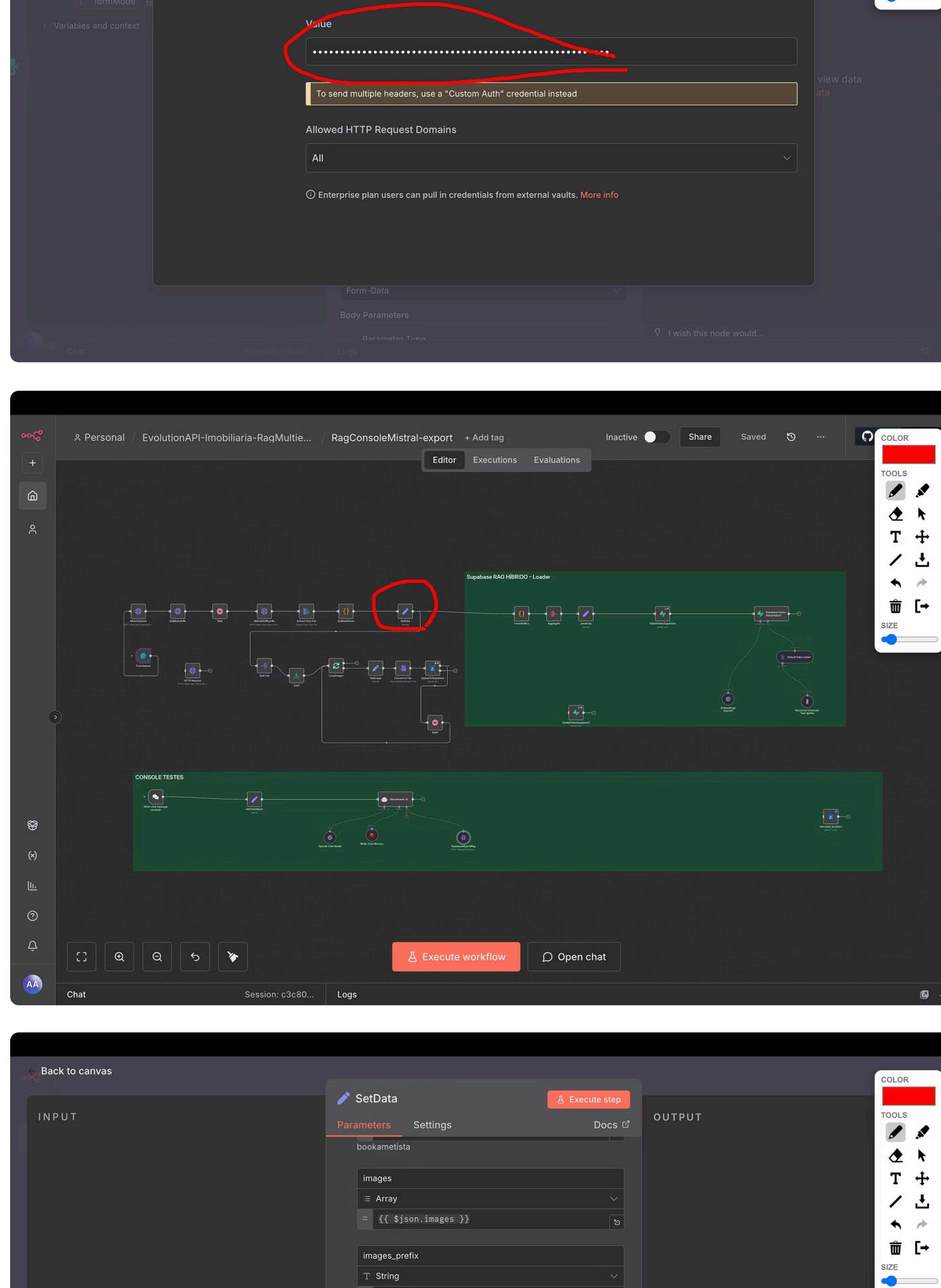
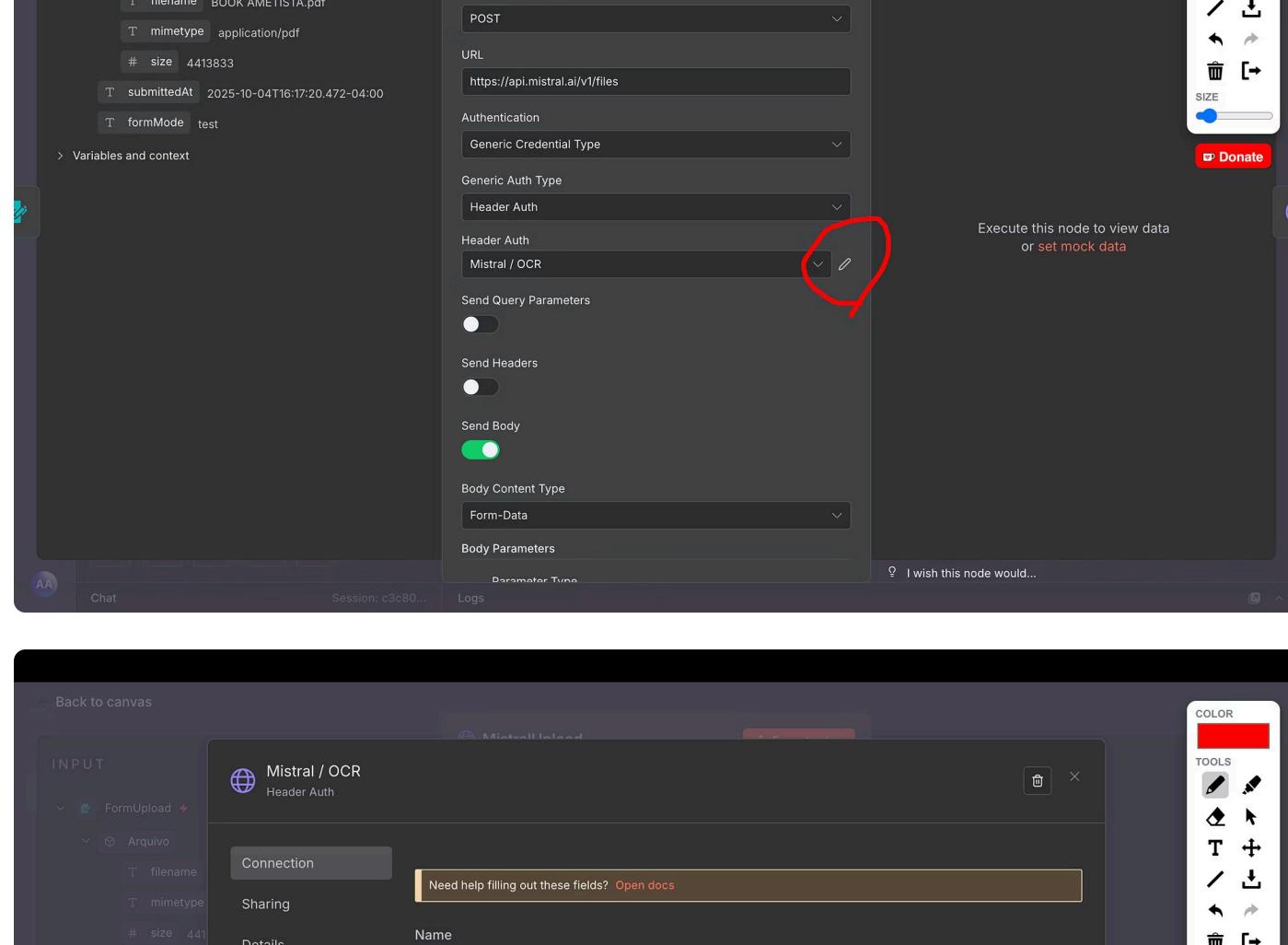
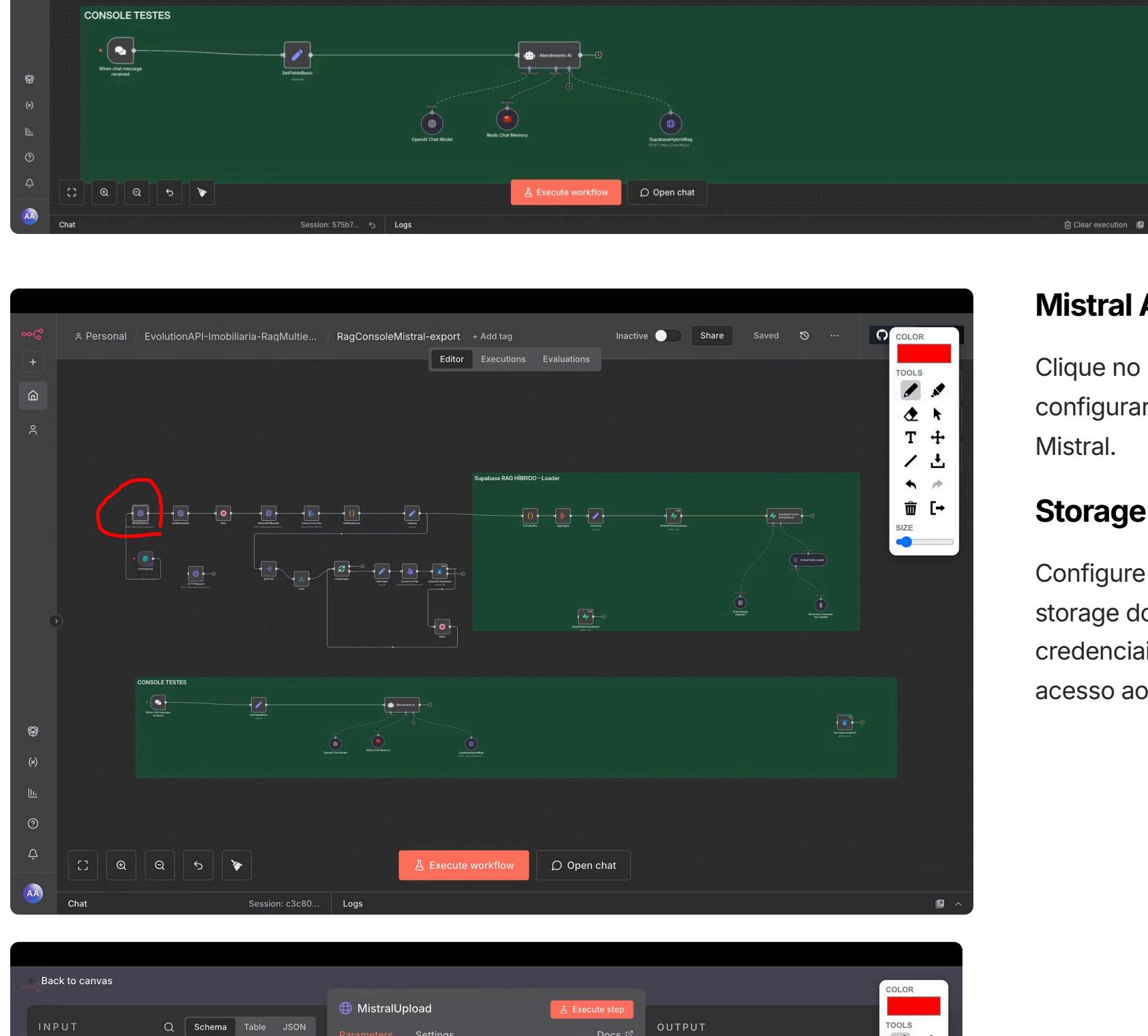
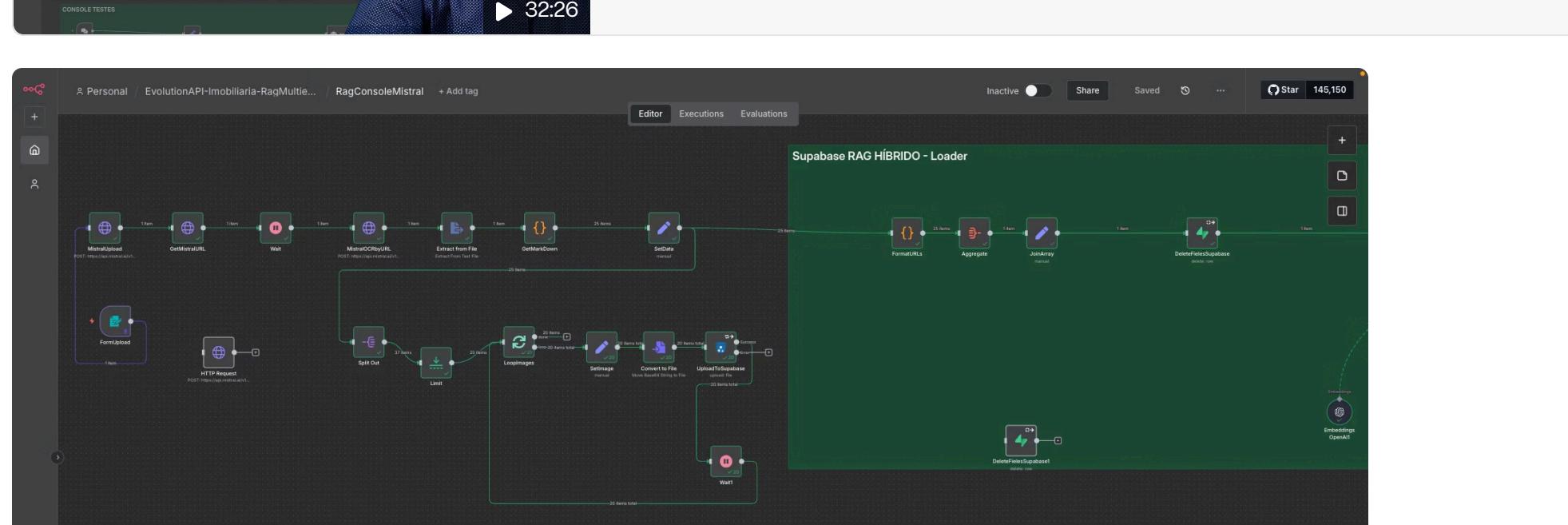
S3 Access Keys

Manage your access keys for this project.

Description	Access key ID	Created at
n8n	AKIAJF66B7AC280F03eb03f7df4fa65f	1 day ago

Fluxo n8n

Este fluxo foi desenvolvido para lidar com documentos complexos — como contratos, relatórios e PDFs com tabelas ou imagens — usando o poder do **Mistral OCR** e a flexibilidade do **n8n**.



Mistral API KEY

Clique no NÓ ao lado para configurar a sua conta da Mistral.

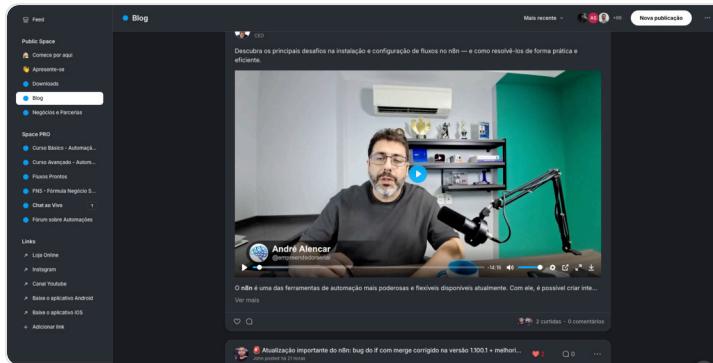
Storage

Configure também a URL do storage do Supabase e credenciais geradas para acesso ao Storage S3

Suporte

Gratuito

Antes de solicitar suporte assista o vídeo gratuito na comunidade do EmpreendedorSerial, aproveite e crie uma conta para ficar por dentro das novidades:



The screenshot shows a social media feed for the 'Blog' section of the EmpreendedorSerial community. A video thumbnail by André Alencar is displayed, titled 'Os Principais Problemas na Instalação e Configuração de Fluxos no n...'. The video description reads: 'Descubra os principais desafios na instalação e configuração de fluxos no n8n — e como resolvê-los de forma prática e eficiente. O n8n é uma das ferramentas ...'. The video has 2 likes and 0 comments.

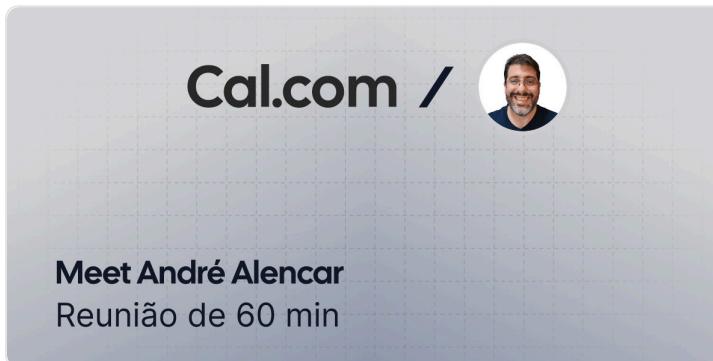
Caso não encontre uma solução envie todos os detalhes para nossos canais de suporte gratuito (dias úteis no horário comercial). Não esqueça de enviar imagens do problema e descrição completa:

- E-mail: suporte@aalencar.com.br
- Whatsapp: **+55 86 9999-7003**

Não esqueça de detalhar seu problema, enviar imagens com os erros ou vídeos relatando o problema.

Pago:

Para suporte pago e consultoria paga escolha um horário disponível em nosso site da cal.com



The screenshot shows a booking interface for Cal.com. It features a profile picture of André Alencar and the text 'Meet André Alencar' followed by 'Reunião de 60 min'. To the right, there is a booking card for 'Reunião de 60 min | André Alencar | Cal.com'.