DevFest Singapore 2023

PaLM API: No-code Prototyping with Flowise

Martin Andrews - Singapore ML GDE

About Me

RED DRAGON AI

- Google Developer Expert for Machine Learning and Deep Learning (2017-)
- Deep Learning R&D :
  - Language & Dialogue systems
  - Generative Models
  - Text-to-Speech
- MeetUp Co-organiser:
  - "Machine Learning Singapore"

Martin Andrews

Google Developers Experts

RED DRAGON AI

About Red Dragon AI

- Founded 2017
- Google Partner
- Consulting, Prototyping & Building
- Research - NeurIPS, EMNLP, COLING, NAACL
- Interactive Digital Personas

Overview

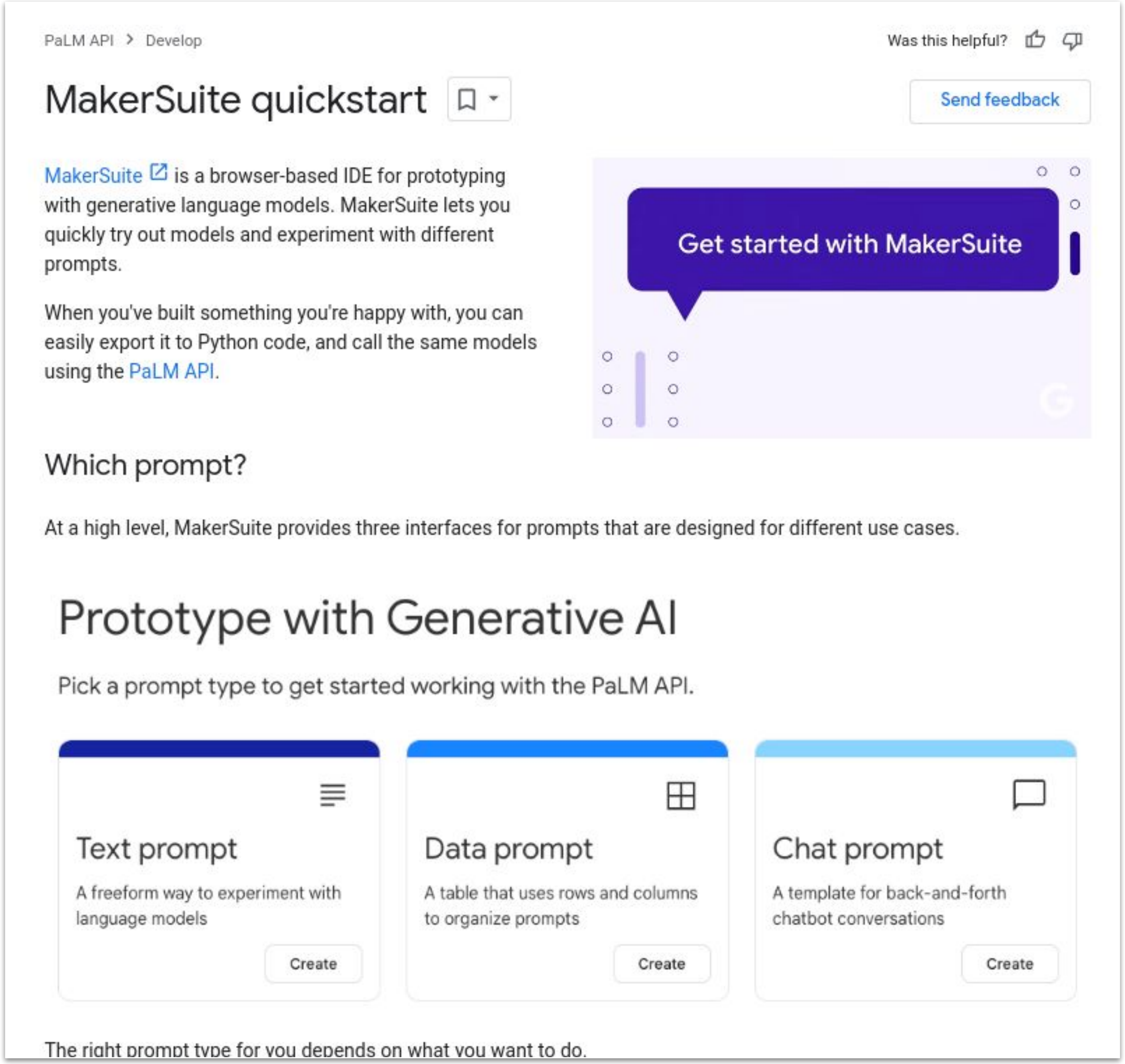
No-code Prototyping with Flowise

- Talk through:
  - MakerSuite PaLM
  - LangChain & Flowise
- Demonstrate
  - <cross-fingers>
    - Simple queries
    - Building a RAG system
  - </cross-fingers>

MakerSuite

What is MakerSuite?

... and how do I get a PaLM API Key?



PaLM API > Develop

Was this helpful?

MakerSuite quickstart

MakerSuite is a browser-based IDE for prototyping with generative language models. MakerSuite lets you quickly try out models and experiment with different prompts.

When you've built something you're happy with, you can easily export it to Python code, and call the same models using the PaLM API.

Which prompt?

At a high level, MakerSuite provides three interfaces for prompts that are designed for different use cases.

Prototype with Generative AI

Pick a prompt type to get started working with the PaLM API.

Text prompt

A freeform way to experiment with language models

Data prompt

A table that uses rows and columns to organize prompts

Chat prompt

A template for back-and-forth chatbot conversations

The right prompt type for you depends on what you want to do.

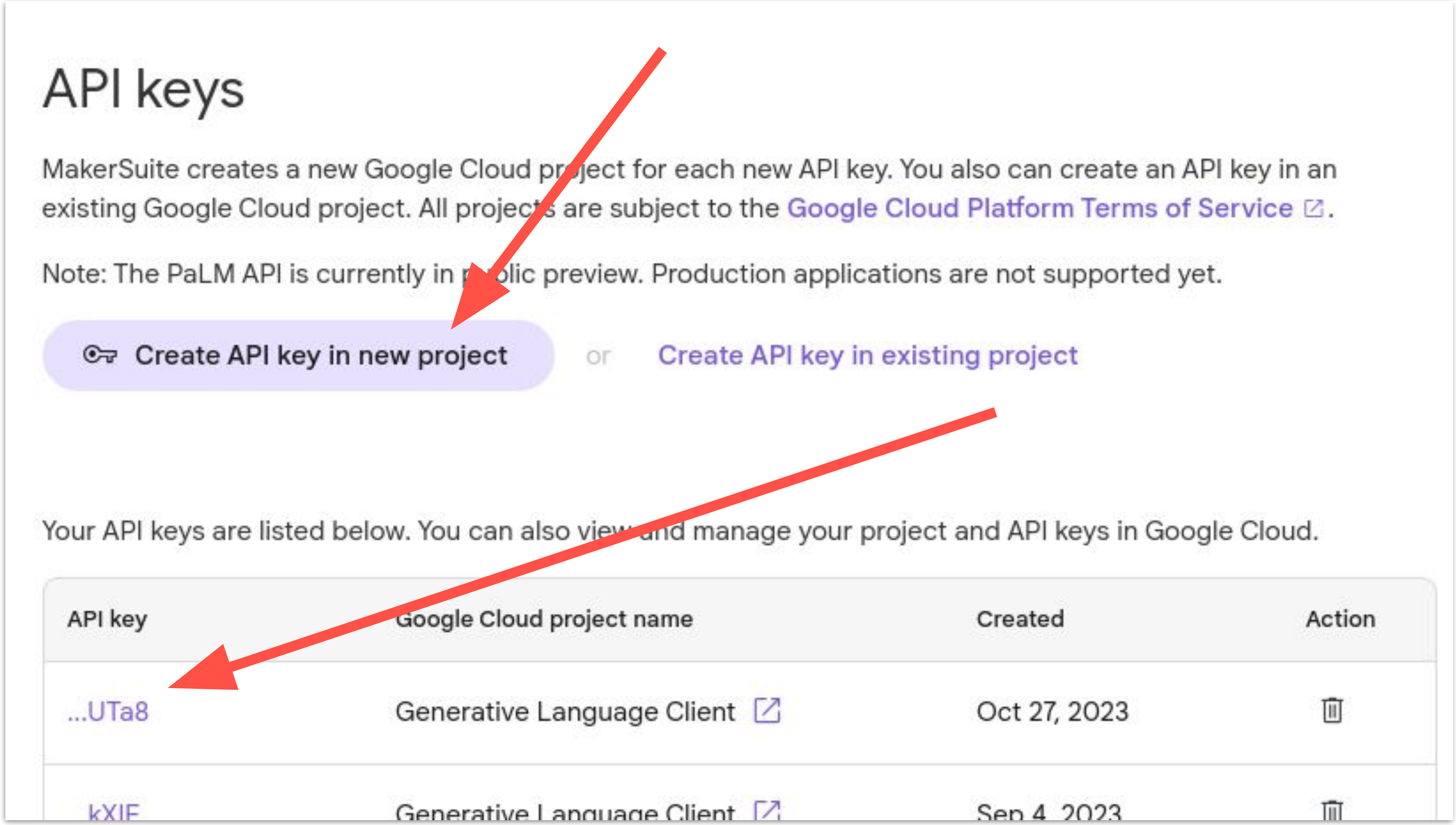
MakerSuite

- "A fast, easy way to start prototyping generative AI ideas"
- ... but doesn't have much flexibility

- Also has API access :
  - PaLM Large Language Model
  - PaLM Chat Model
  - PaLM embeddings

PaLM API

Getting an API Key



API keys

MakerSuite creates a new Google Cloud project for each new API key. You also can create an API key in an existing Google Cloud project. All projects are subject to the Google Cloud Platform Terms of Service.

Note: The PaLM API is currently in public preview. Production applications are not supported yet.

Create API key in new project
or
Create API key in existing project

Your API keys are listed below. You can also view and manage your project and API keys in Google Cloud.

API key	Google Cloud project name	Created	Action
...UTa8	Generative Language Client	Oct 27, 2023	🗑
...kXIE	Generative Language Client	Sep 4, 2023	🗑

https:// makersuite.google.com / app / apikey

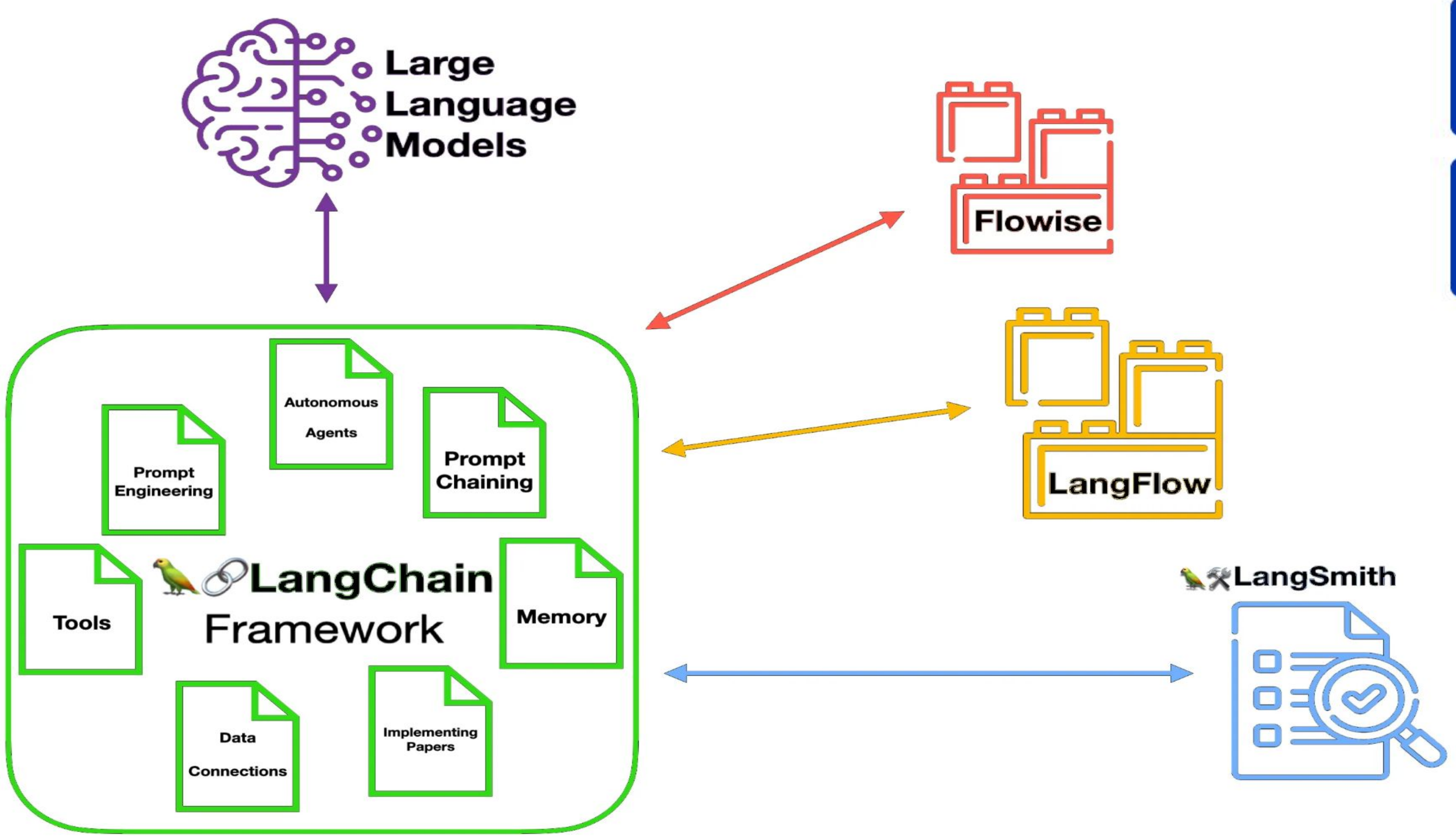
langchain

What is langchain?

… and how does Flowise fit in?

langchain

The LangChain Ecosystem



www.cobusgreyling.com

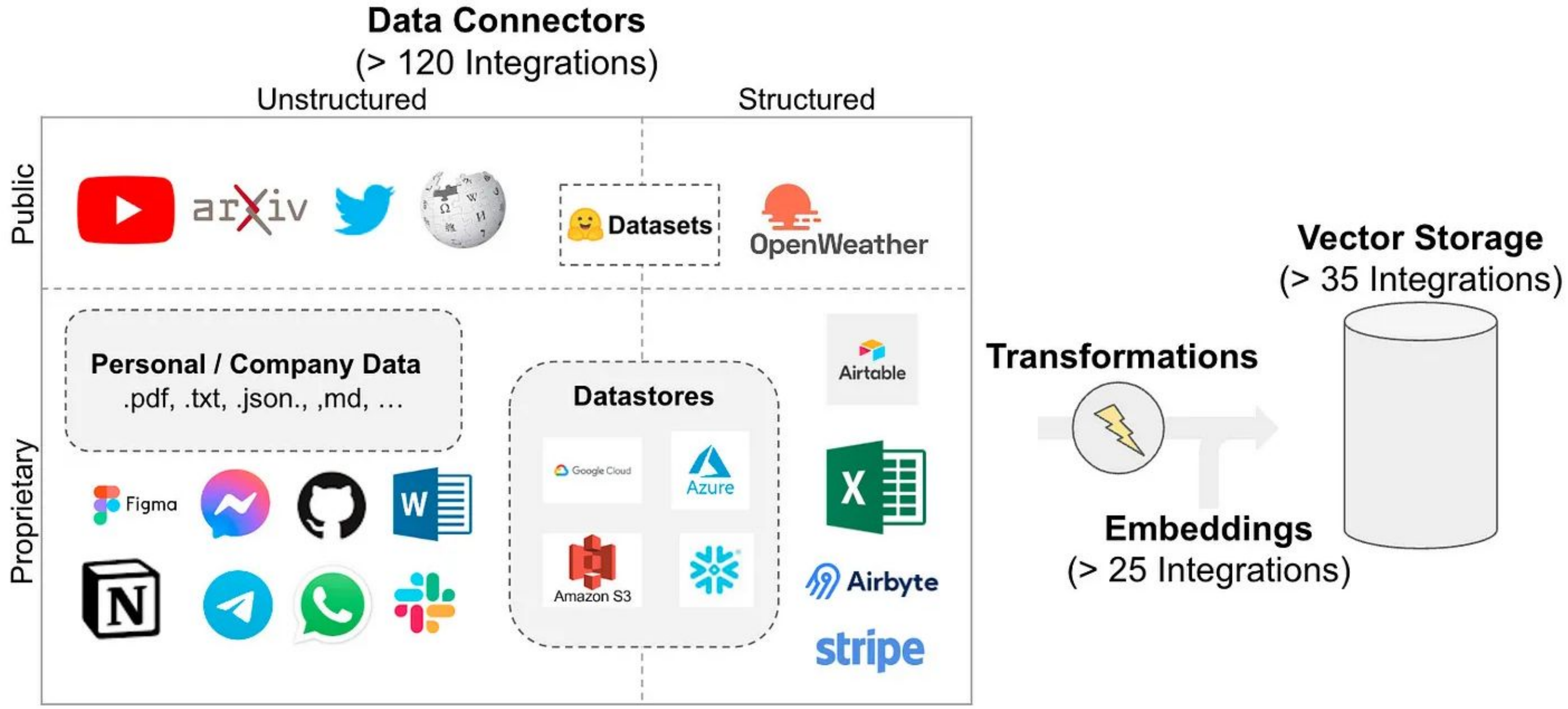
5.3M
Monthly Downloads

67k
Github Stars

1800+
Contributors

34k+
Apps Powered

langchain



Data Interfaces

langchain

Using LangChain

- LangChain is very flexible
  - And growing very quickly
  - New ideas implemented continuously

- Primary users = Power users
  - Programming in Python

Flowise

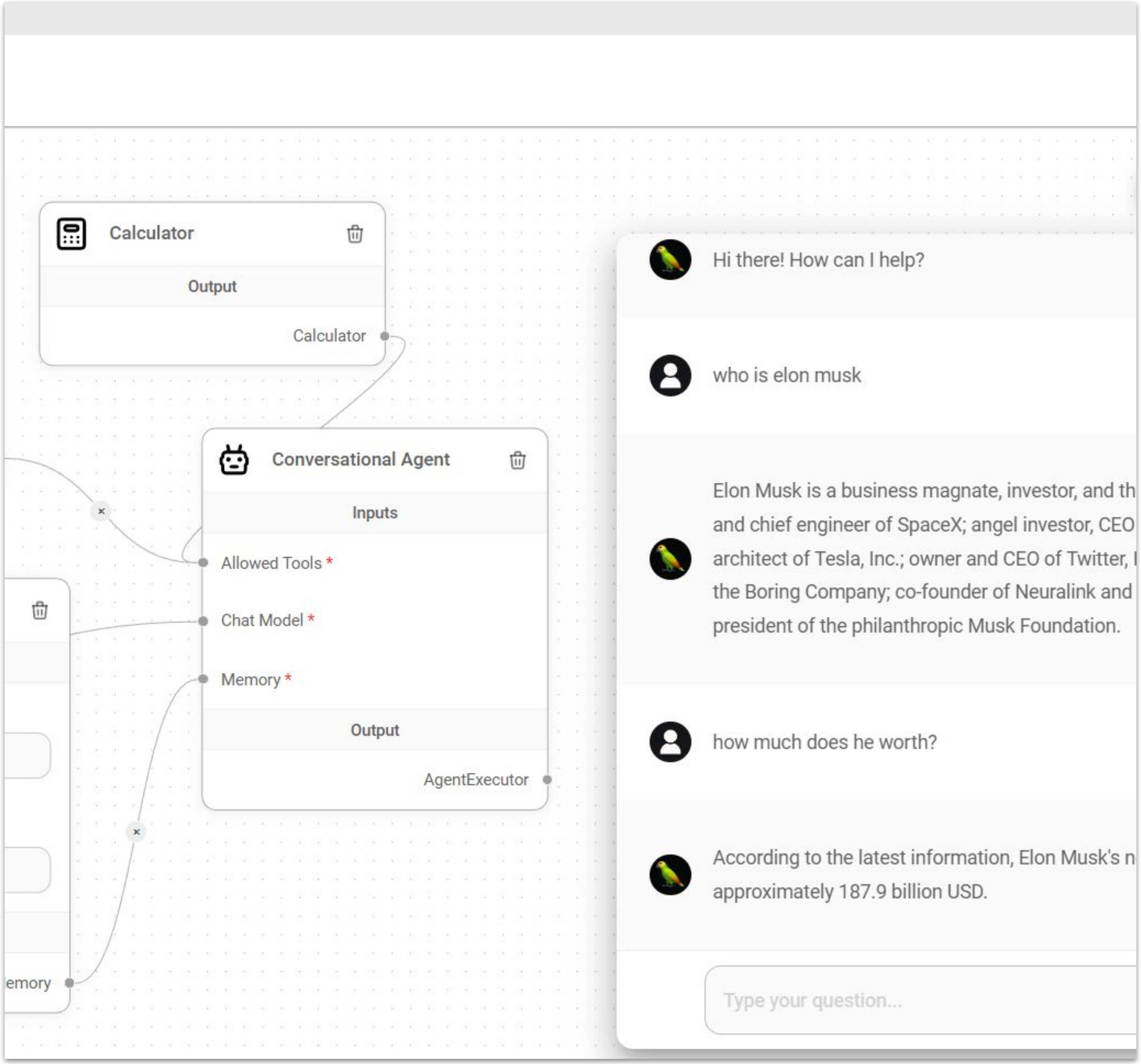
What is Flowise?

... and how is it used?

Flowise

Key Ideas

- Need a web-based GUI to make LangChain more accessible
- Flowise is Javascript/Typescript
  - Instead of the Python backend, uses `langchain.js`
- Front-end UI requires no coding
  - Still have access to ecosystem



Hi there! How can I help?

who is elon musk

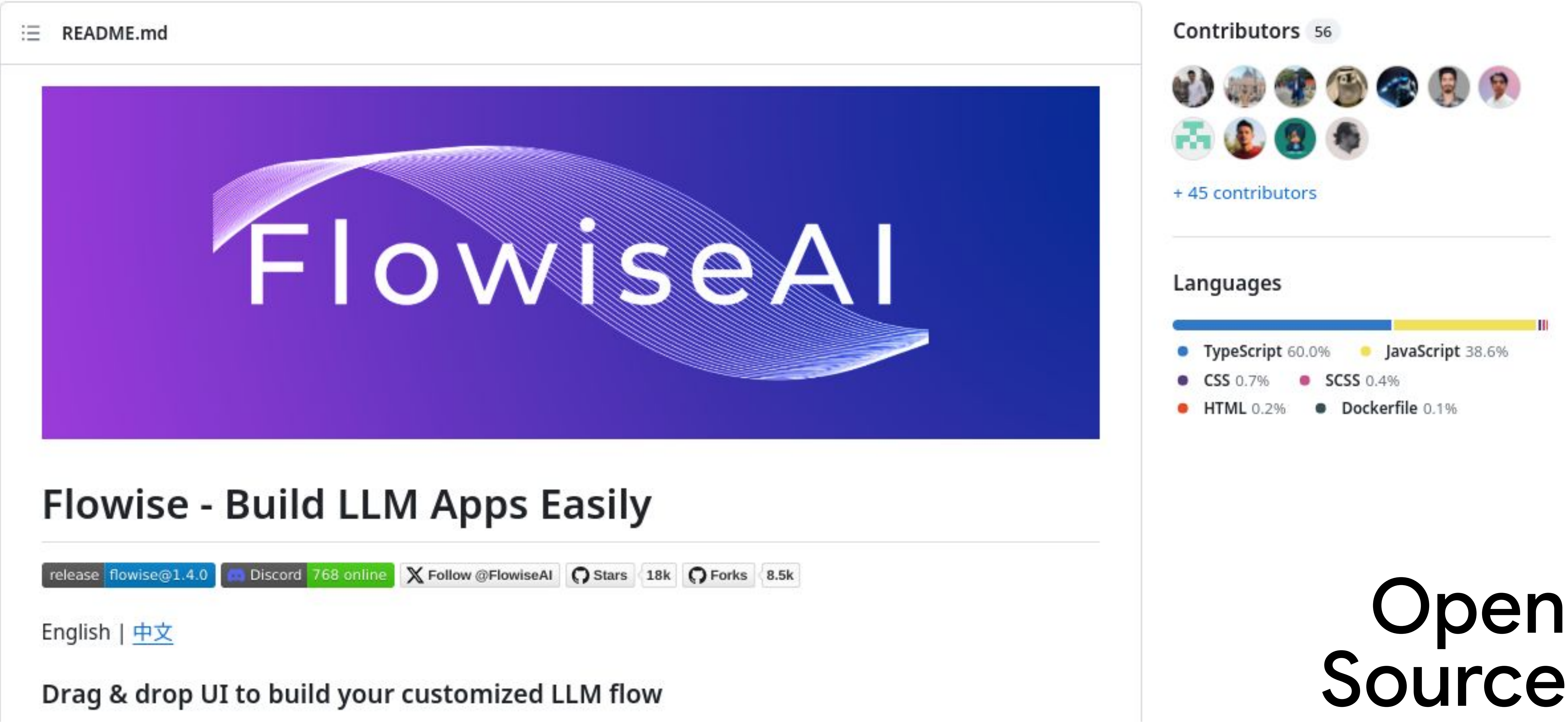
Elon Musk is a business magnate, investor, and th
and chief engineer of SpaceX; angel investor, CEO
architect of Tesla, Inc.; owner and CEO of Twitter, I
the Boring Company; co-founder of Neuralink and
president of the philanthropic Musk Foundation.

how much does he worth?

According to the latest information, Elon Musk's n
approximately 187.9 billion USD.

Type your question...

Flowise



README.md

Flowise - Build LLM Apps Easily

release flowise@1.4.0 | Discord 768 online | Follow @FlowiseAI | Stars 18k | Forks 8.5k

English | 中文

Drag & drop UI to build your customized LLM flow

Contributors 56

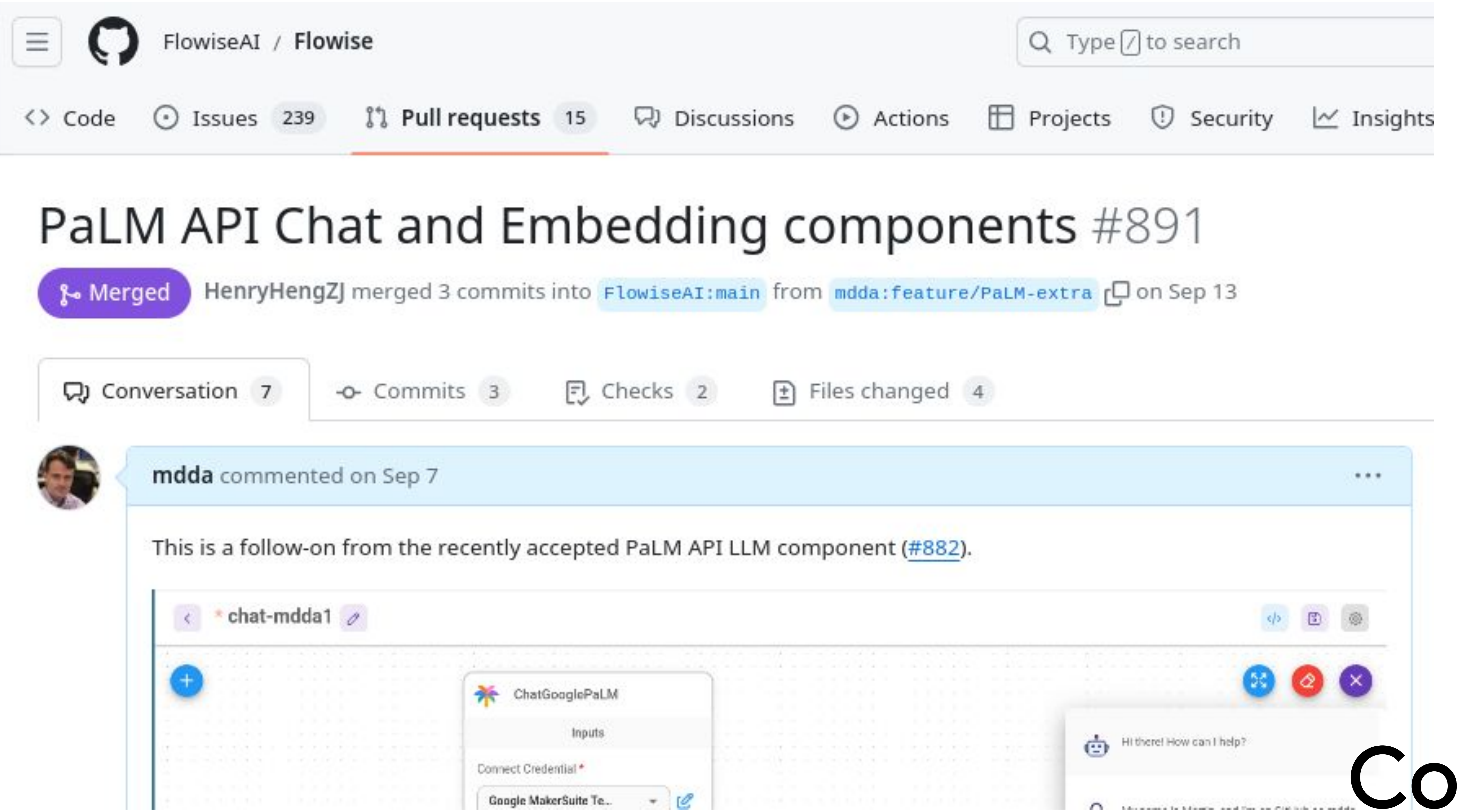
+ 45 contributors

Languages

● TypeScript 60.0%	● JavaScript 38.6%
● CSS 0.7%	● SCSS 0.4%
● HTML 0.2%	● Dockerfile 0.1%

Open
Source

Flowise



PaLM API Chat and Embedding components #891

Merged HenryHengZJ merged 3 commits into `FlowiseAI:main` from `mdda:feature/PaLM-extra` on Sep 13

mdda commented on Sep 7

This is a follow-on from the recently accepted PaLM API LLM component (#882).

Hi there! How can I help?

My Contribution

Flowise

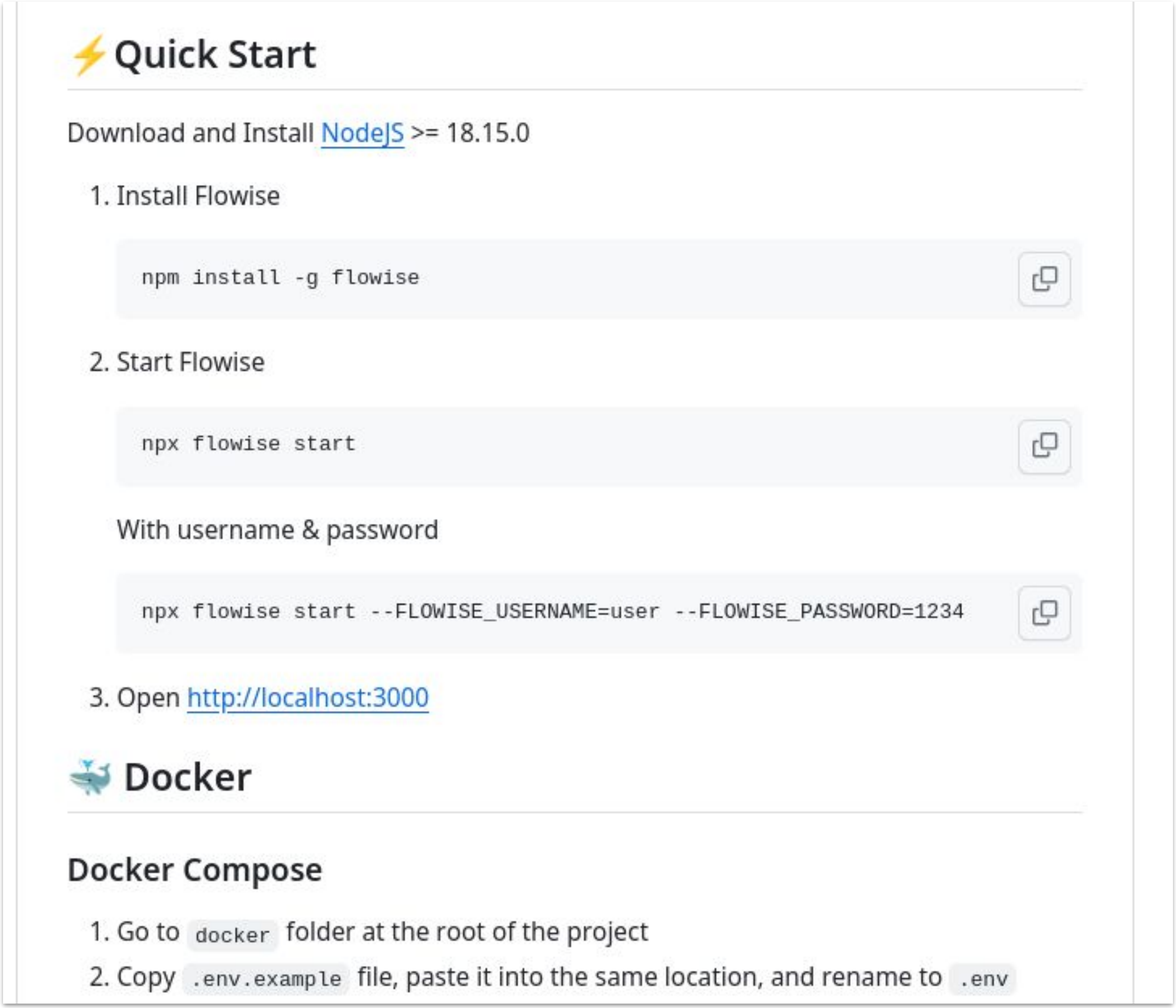
Demo Time!

< cross-fingers > ...

Flowise

Launch the Backend

- (On GCP)



⚡Quick Start

Download and Install NodeJS >= 18.15.0

1. Install Flowise

```
npm install -g flowise
```

2. Start Flowise

```
npx flowise start
```

With username & password

```
npx flowise start --FLOWISE_USERNAME=user --FLOWISE_PASSWORD=1234
```

3. Open http://localhost:3000

🐳 Docker

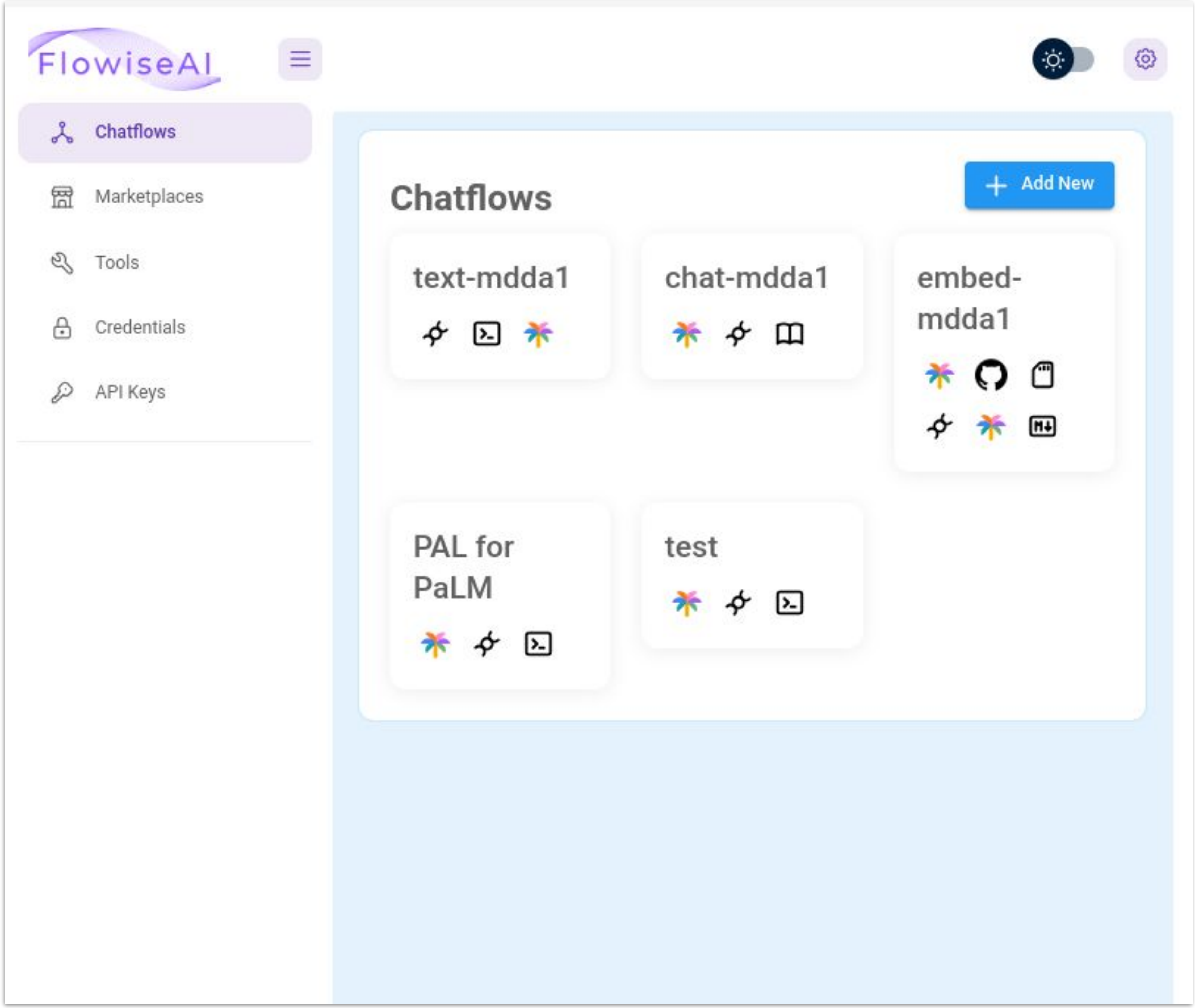
Docker Compose

1. Go to `docker` folder at the root of the project
2. Copy `.env.example` file, paste it into the same location, and rename to `.env`

Flowise

Launch the User Interface

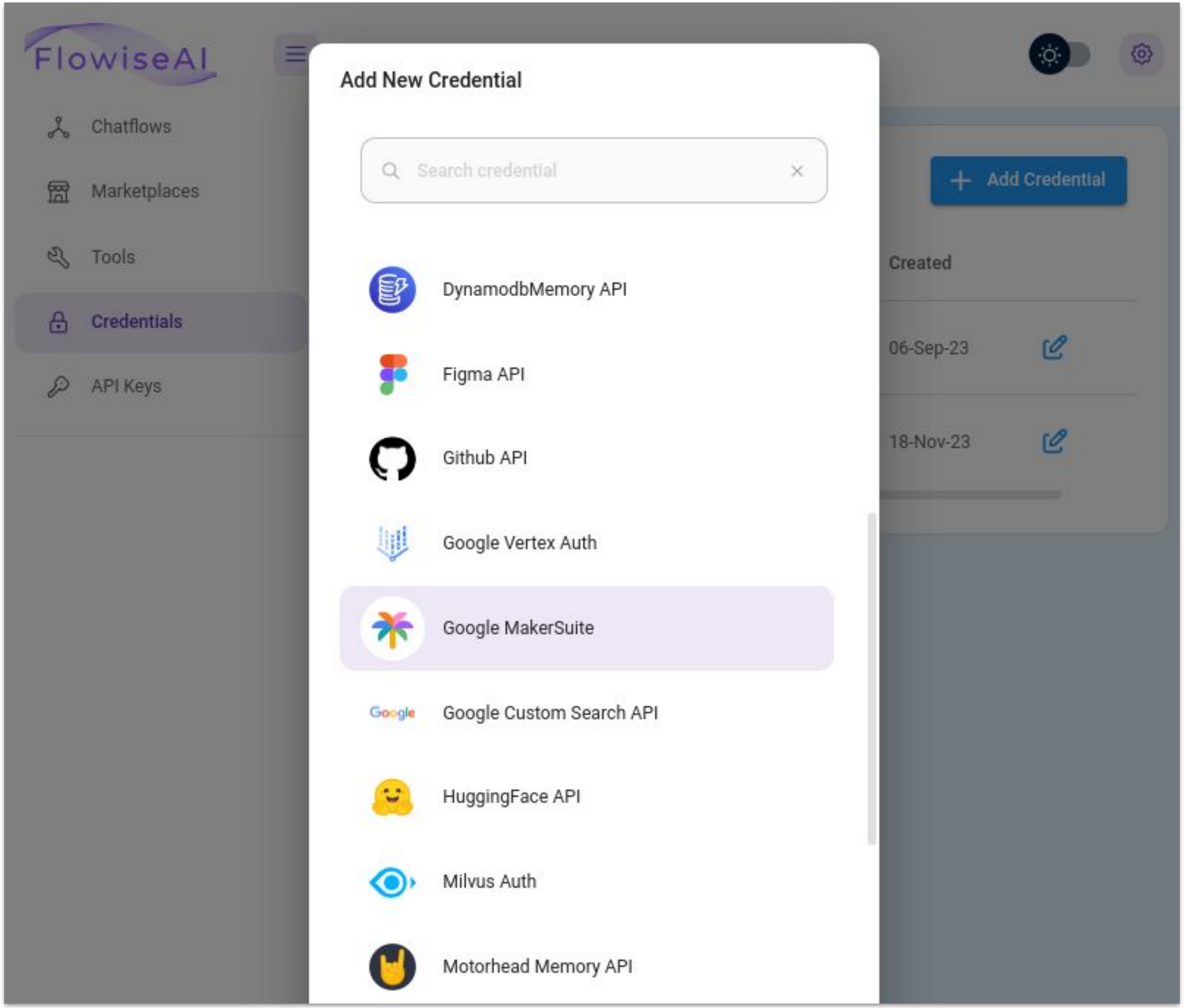
- (In Browser)



Flowise

Add in PaLM API credentials

- Get a PaLM API Key from the MakerSuite site
- (add key into Flowise store)



Add New Credential

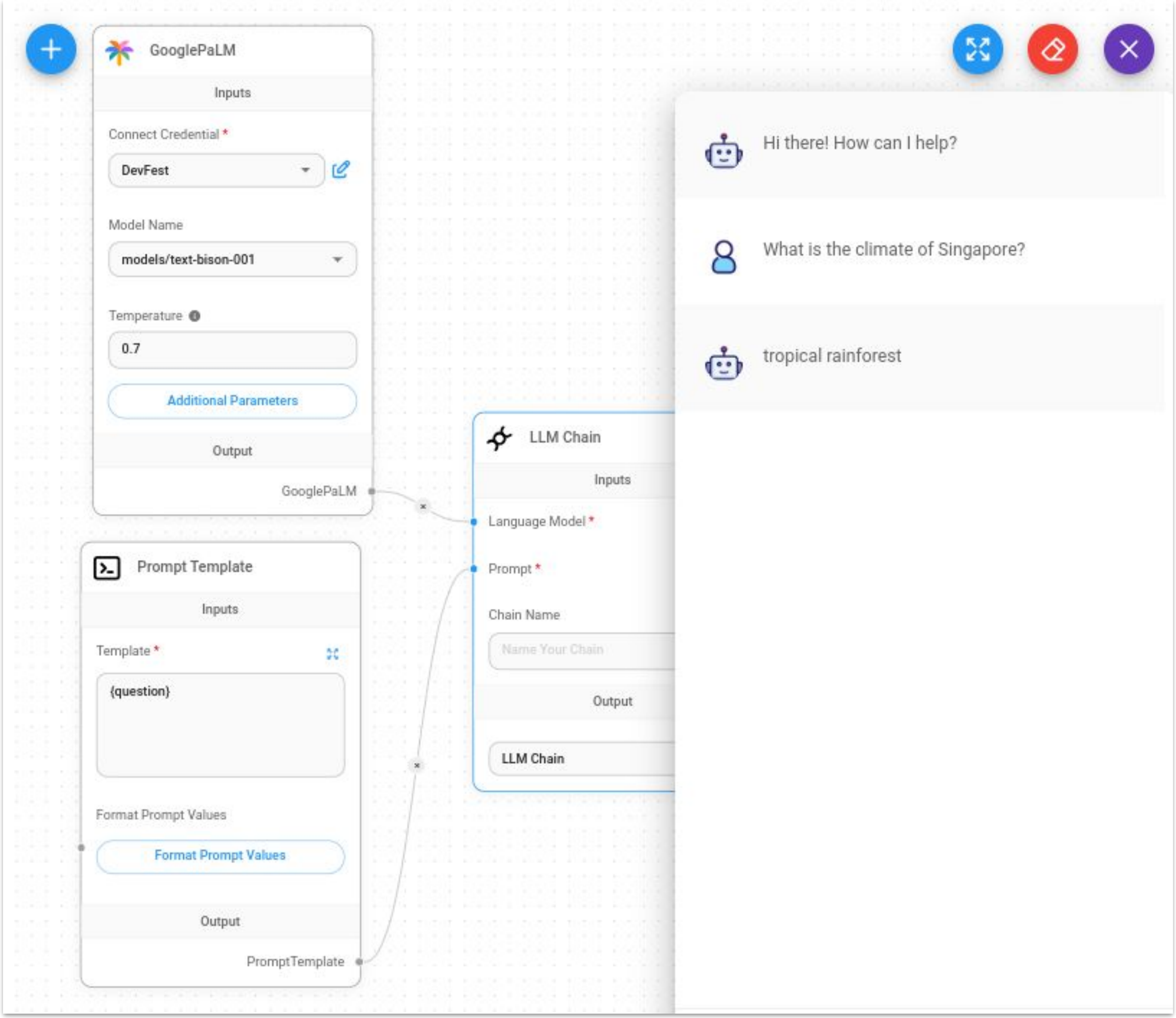
Search credential

- DynamodbMemory API
- Figma API
- Github API
- Google Vertex Auth
- Google MakerSuite
- Google Custom Search API
- HuggingFace API
- Milvus Auth
- Motorhead Memory API

Flowise

First query of the PaLM LLM

- What is the climate of Singapore?



Hi there! How can I help?

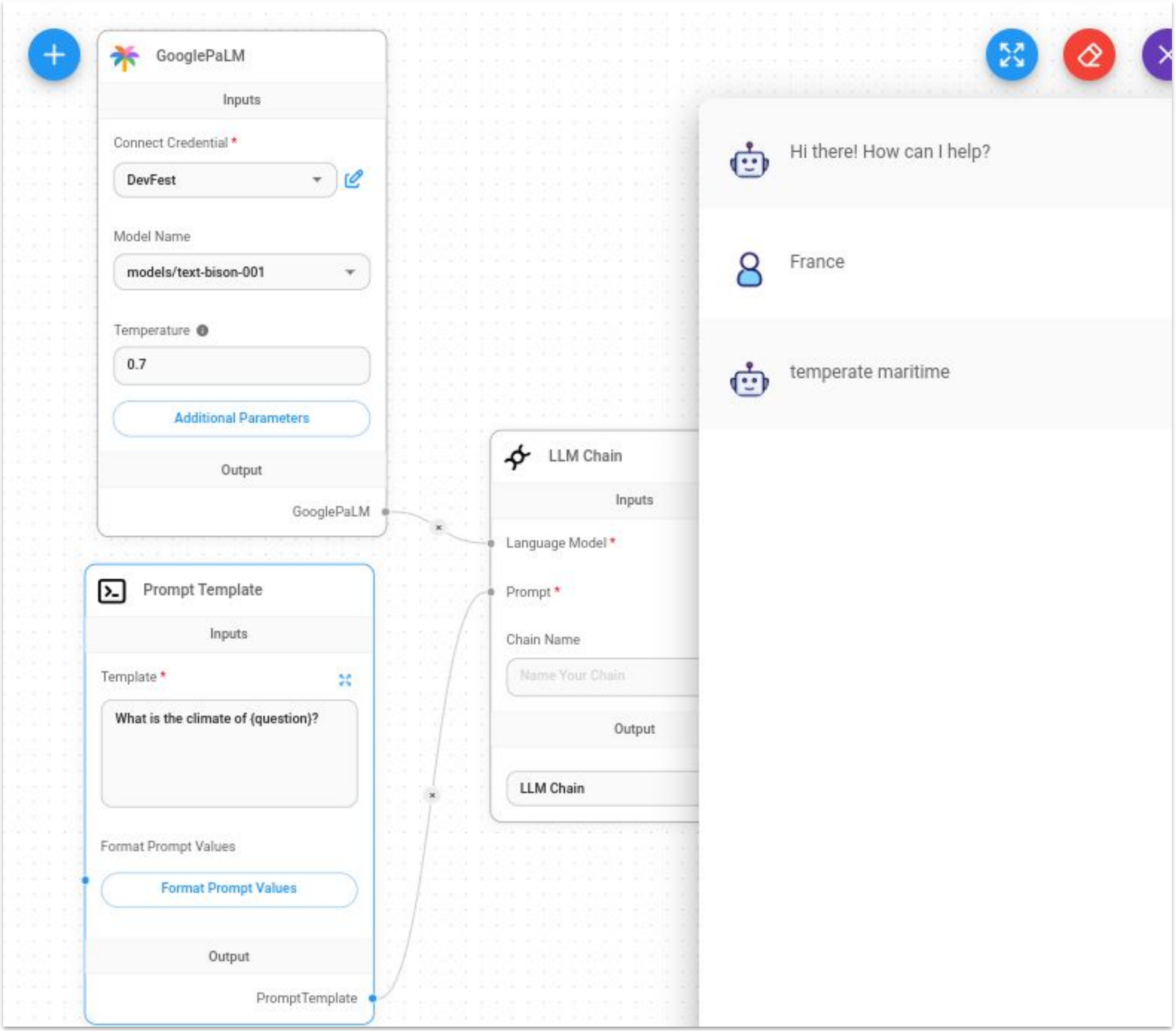
What is the climate of Singapore?

tropical rainforest

Flowise

Create a Templated Response

- What is the climate of {X} ?



GooglePaLM

Connect Credential *: DevFest

Model Name: models/text-bison-001

Temperature: 0.7

Prompt Template

Template *: What is the climate of {question}?

LLM Chain

Chain Name: Name Your Chain

Hi there! How can I help?

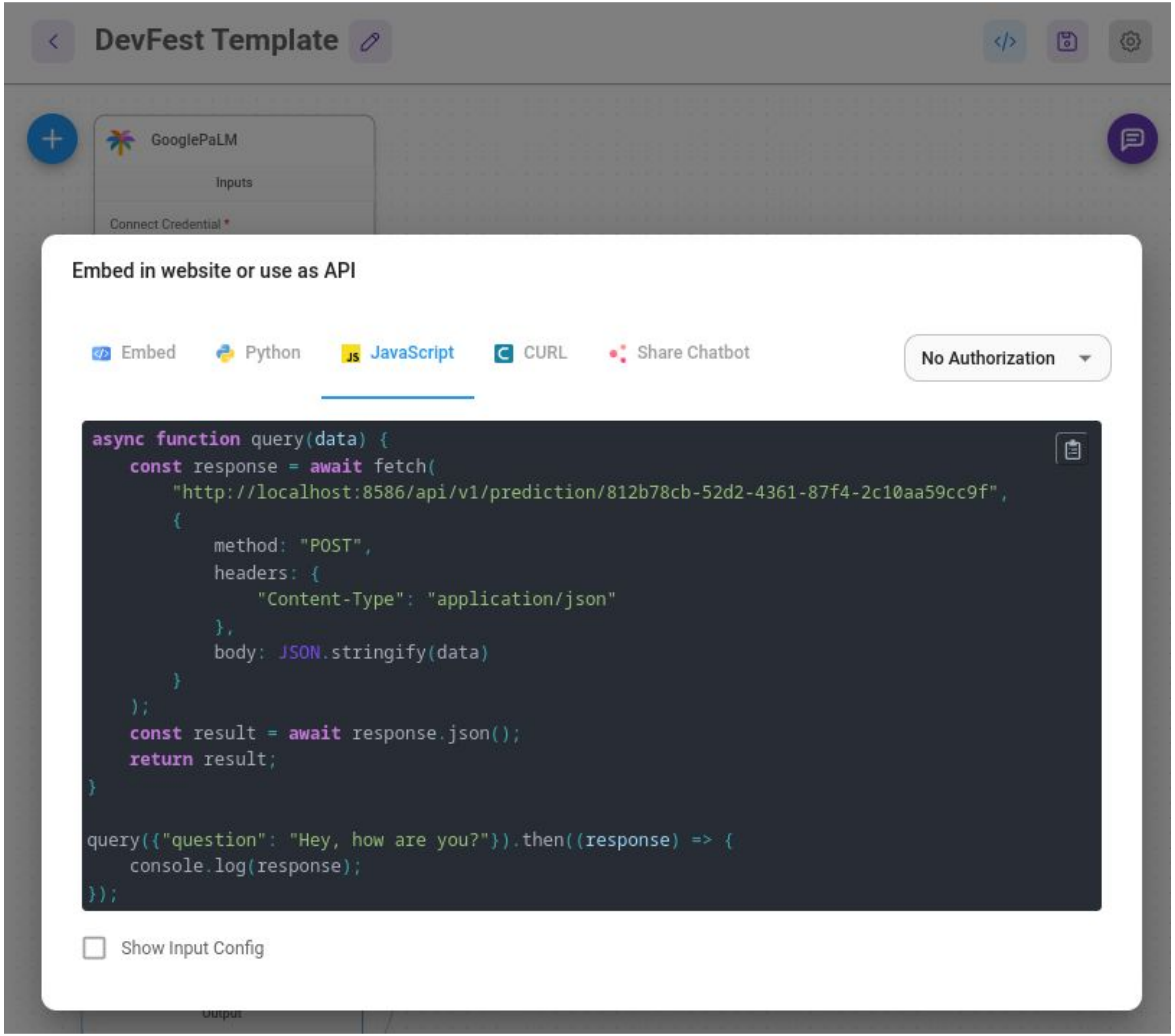
France

temperate maritime

Flowise

Instance REST API

- All of our UI plans automatically create equivalent REST APIs



```
async function query(data) {
    const response = await fetch(
        "http://localhost:8586/api/v1/prediction/812b78cb-52d2-4361-87f4-2c10aa59cc9f",
        {
            method: "POST",
            headers: {
                "Content-Type": "application/json"
            },
            body: JSON.stringify(data)
        }
    );
    const result = await response.json();
    return result;
}

query({"question": "Hey, how are you?"}).then((response) => {
    console.log(response);
});
```

☐ Show Input Config

Flowise : RAG

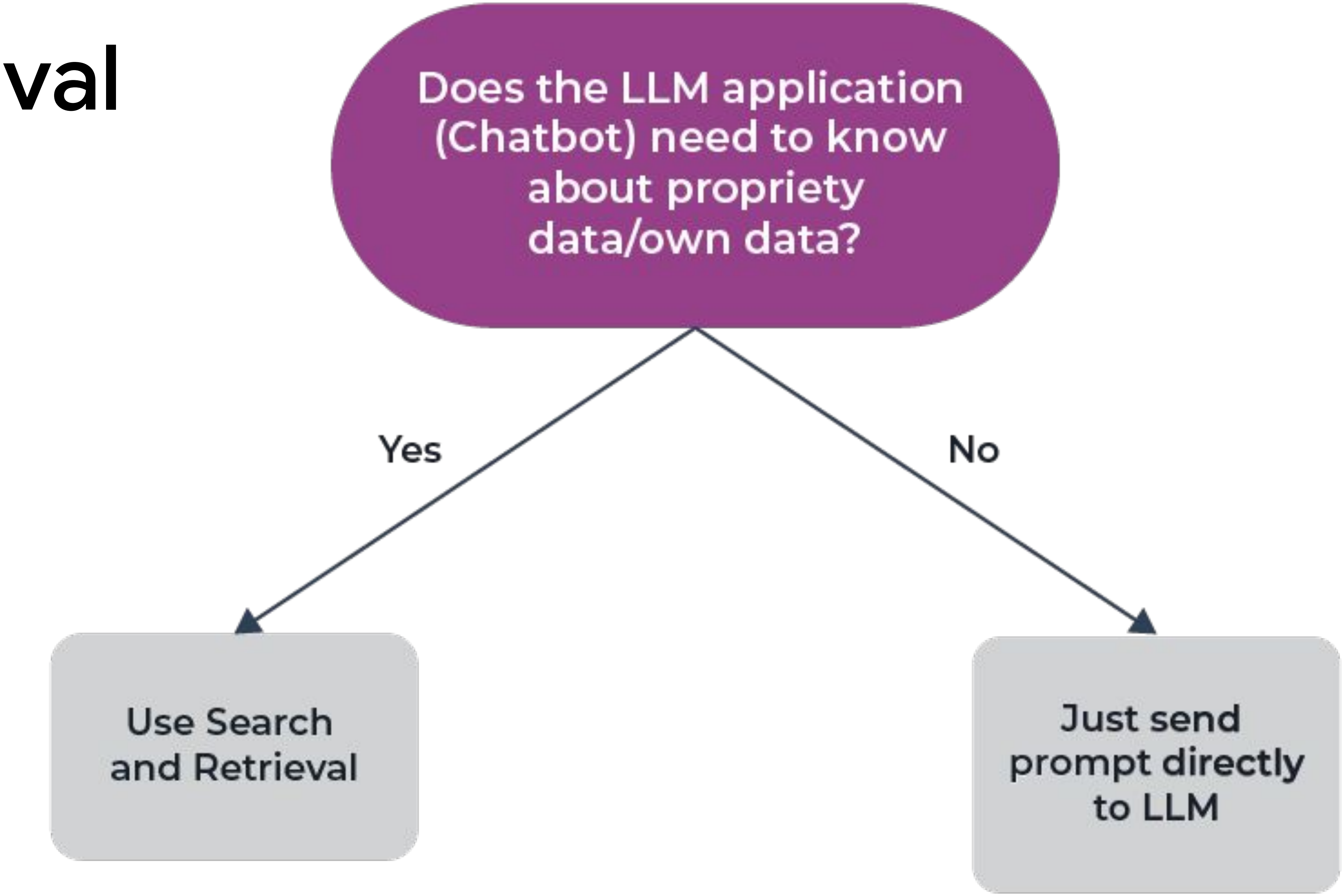
Advanced Demo Time!

Hope for a miracle :-)

Flowise : RAG

Why use Retrieval Augmented Generation (RAG)?

- Key issues:
  - Reduce fabrications
  - Private/specific data
  - Updatable knowledge


Does the LLM application (Chatbot) need to know about propriety data/own data?
Yes
No
Use Search and Retrieval
Just send prompt directly to LLM

Flowise : RAG

Elements of RAG

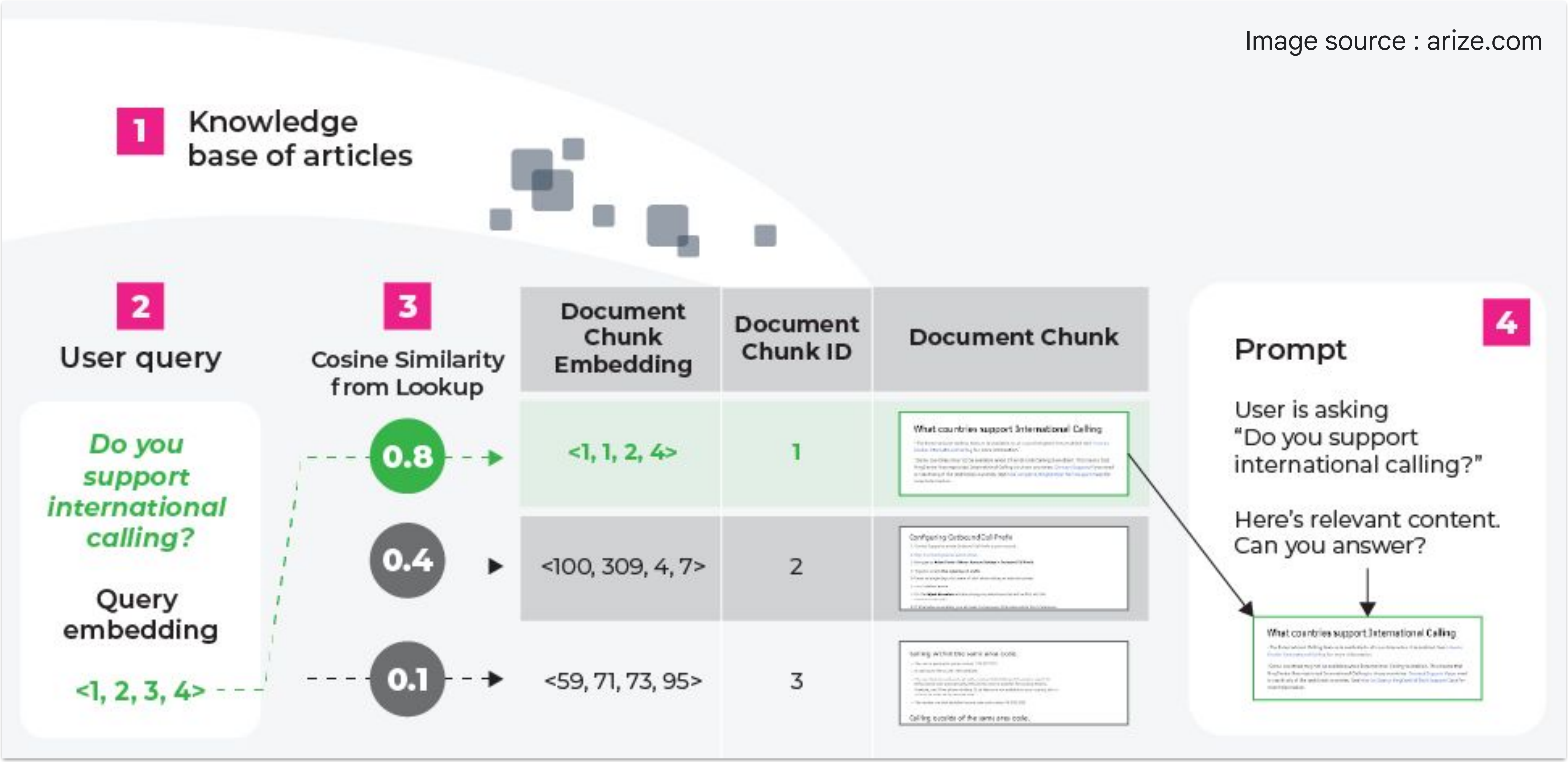


Image source : arize.com

1 Knowledge base of articles

2 User query

Do you support international calling?

Query embedding

<1, 2, 3, 4>

3 Cosine Similarity from Lookup

Cosine Similarity from Lookup	Document Chunk Embedding	Document Chunk ID	Document Chunk
0.8	<1, 1, 2, 4>	1	What countries support International Calling
0.4	<100, 309, 4, 7>	2	Configuring Outbound Call Profile
0.1	<59, 71, 73, 95>	3	Calling within the same area code Calling outside of the same area code.

4 Prompt

User is asking "Do you support international calling?"

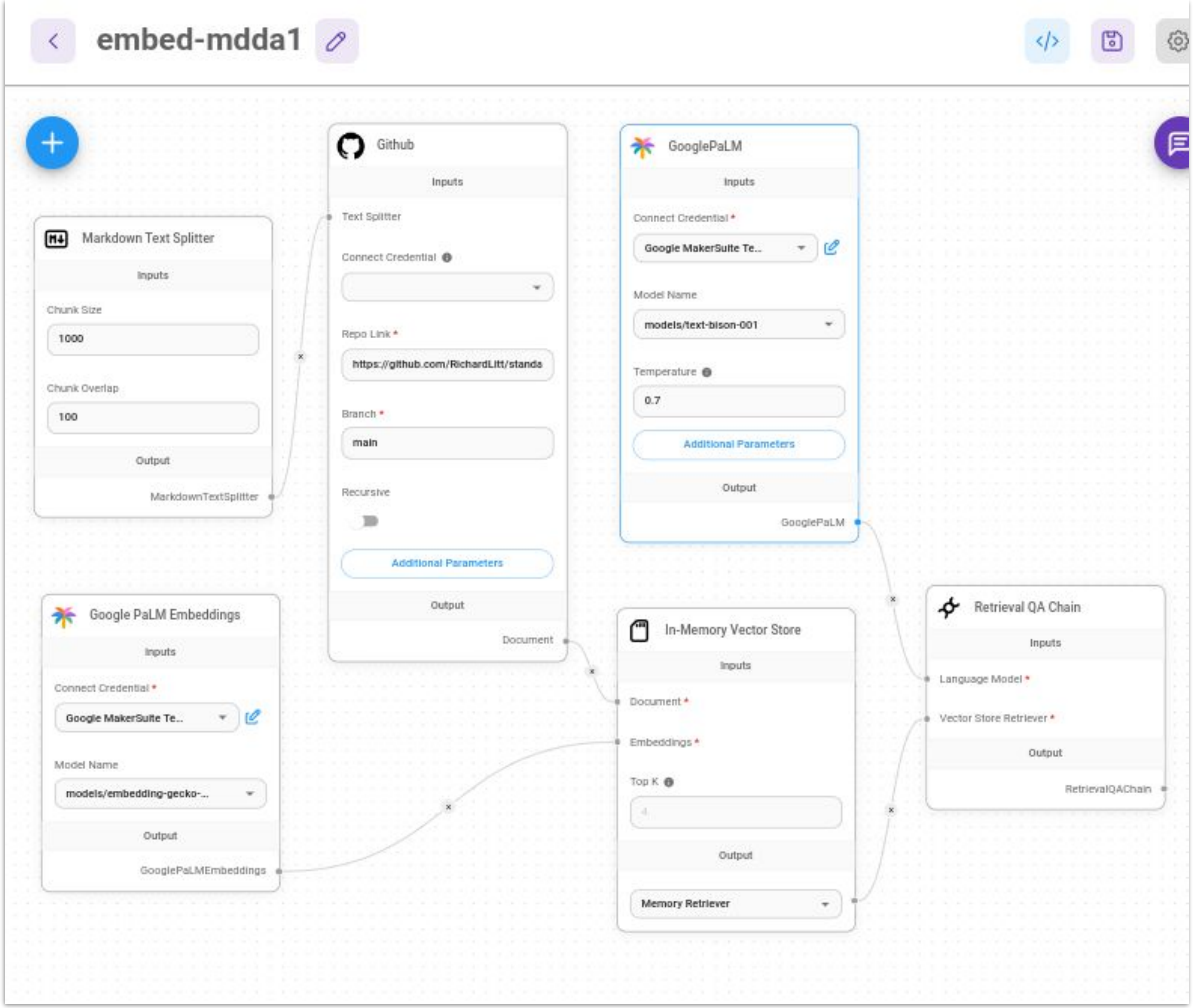
Here's relevant content. Can you answer?

What countries support International Calling

Flowise : RAG

Create a RAG System !

- Steps:
  - Data source
  - Data splitter
  - Embeddings
  - Vector Store
  - LLM for Q&A



embed-mdda1

Markdown Text Splitter
Chunk Size: 1000
Chunk Overlap: 100

Github
Text Splitter
Connect Credential
Repo Link *: https://github.com/RichardLitt/standa
Branch *: main
Recursive

GooglePaLM
Connect Credential *: Google MakerSuite Te..
Model Name: models/text-bison-001
Temperature: 0.7

Google PaLM Embeddings
Connect Credential *: Google MakerSuite Te..
Model Name: models/embedding-gecko-...

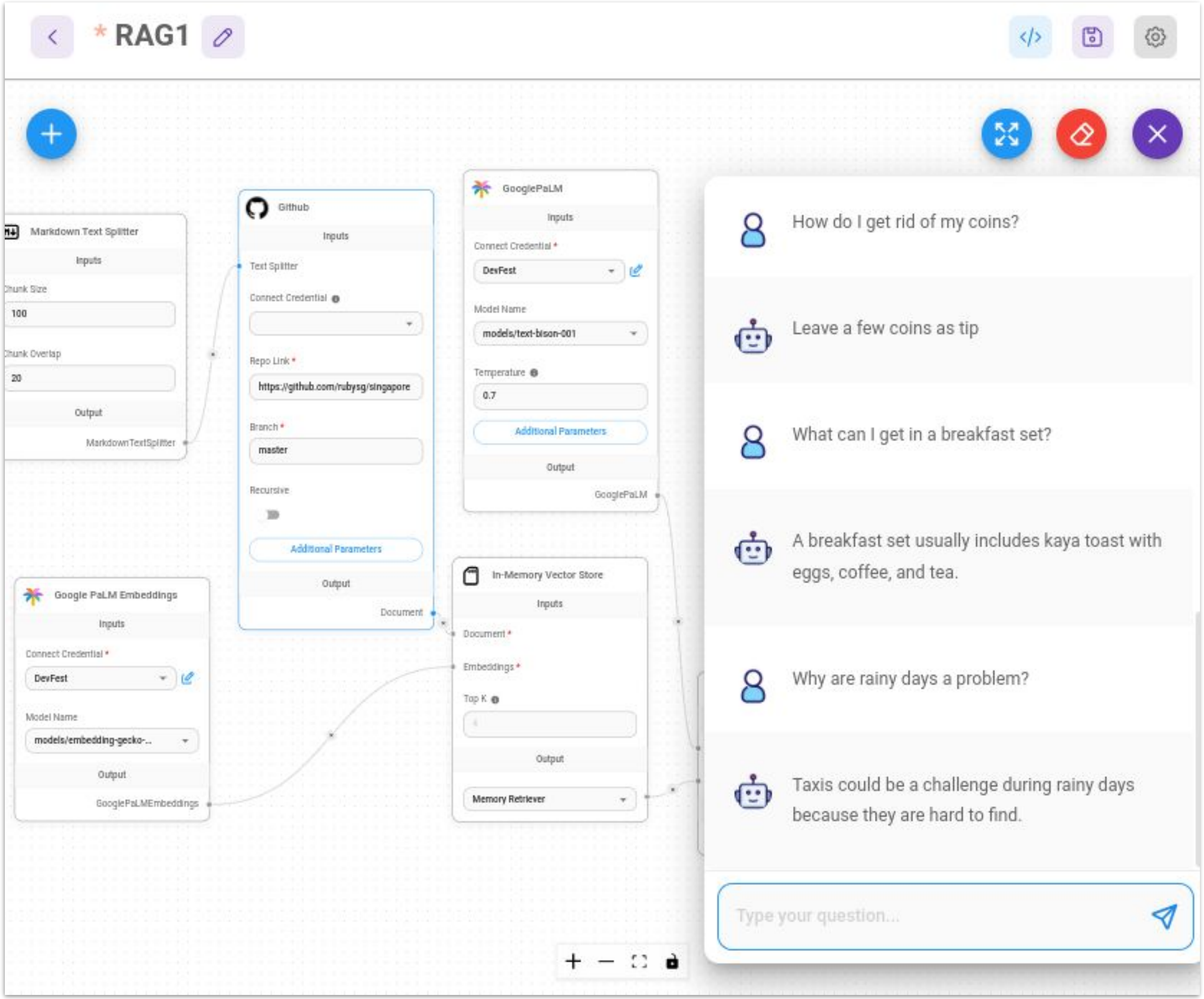
In-Memory Vector Store
Document *
Embeddings *
Top K: 4
Output: Memory Retriever

Retrieval QA Chain
Language Model *
Vector Store Retriever *

Flowise : RAG

Create a RAG System for SG!

- Steps:
  - Data source (master)
  - Data splitter (100 : 20)
  - PaLM Embeddings
  - In-memory Vector Store
  - PaLM LLM for Retrieval Q&A Chain



Markdown Text Splitter
Inputs
Chunk Size: 100
Chunk Overlap: 20
Output
MarkdownTextSplitter

Github
Inputs
Text Splitter
Connect Credential
Repo Link *: https://github.com/rubysg/singapore
Branch *: master
Recursive
Additional Parameters
Output
Document

GooglePaLM
Inputs
Connect Credential *: DevFest
Model Name: models/text-bison-001
Temperature: 0.7
Additional Parameters
Output
GooglePaLM

Google PaLM Embeddings
Inputs
Connect Credential *: DevFest
Model Name: models/embedding-gecko-...
Output
GooglePaLMEmbeddings

In-Memory Vector Store
Inputs
Document *
Embeddings *
Top K
Output
Memory Retriever

How do I get rid of my coins?

Leave a few coins as tip

What can I get in a breakfast set?

A breakfast set usually includes kaya toast with eggs, coffee, and tea.

Why are rainy days a problem?

Taxis could be a challenge during rainy days because they are hard to find.

Type your question...

Flowise

Wrap Up

< / cross-fingers >

Wrap-Up

Low-Code Ready?

Positives :

- UI for langchain very helpful
- Can achieve PoC directly
- ... and immediately have API

Negatives :

- UI can lag behind main langchain
- Still need to understand the issues that LLMs have - and how to fix them

ML Singapore MeetUp

Machine Learning Singapore MeetUp

- Our MeetUp has over 5,600 members now! : Monthly in-person meetings
- Go to : https://www.meetup.com/machine-learning-singapore/

Google I/O Extended

Thank you!

Martin Andrews

Head of AI : Red Dragon AI
Google Developer Expert Machine Learning & Deep Learning

@mdda123 @mdda