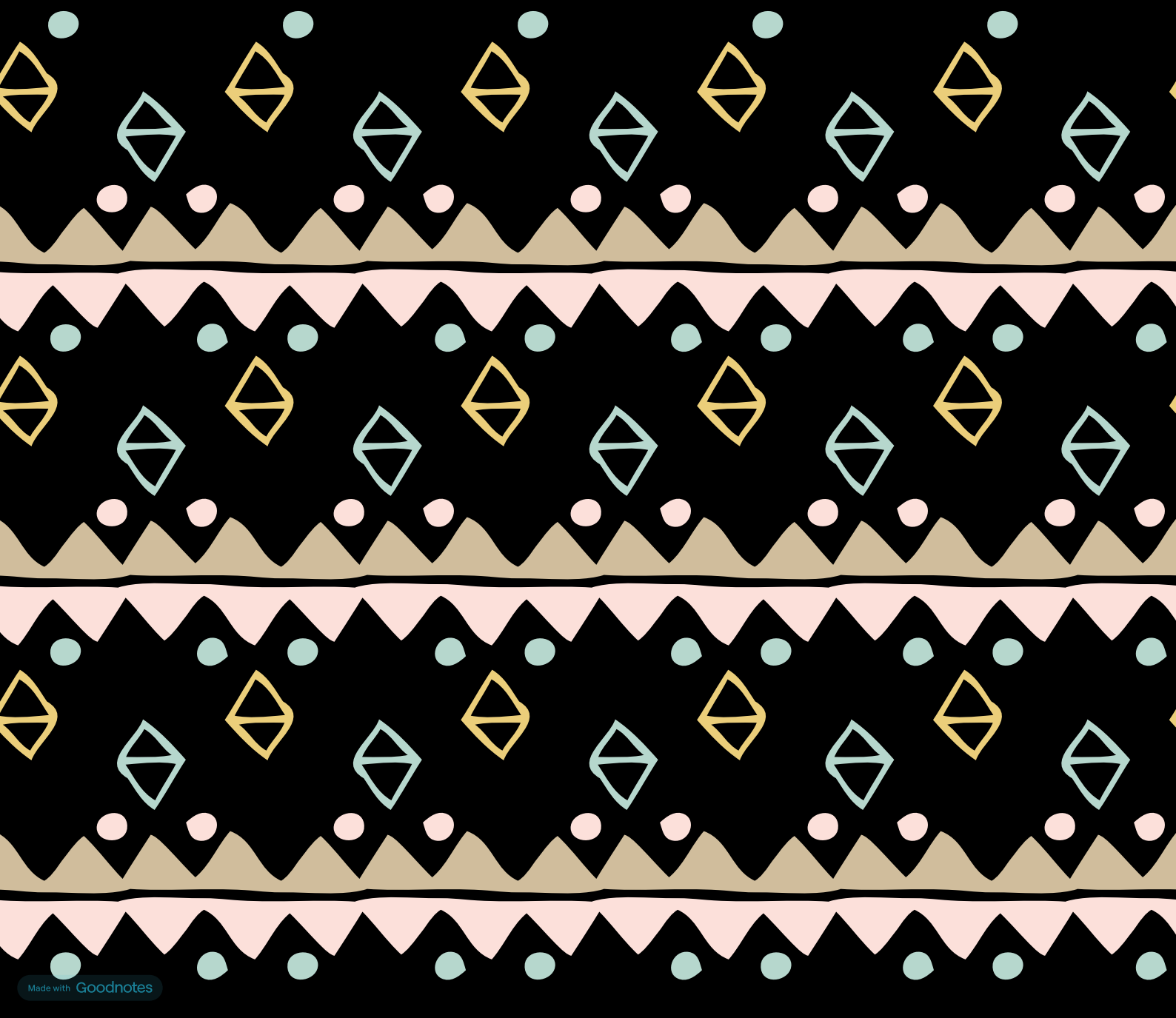
Riddhi Airan

AI Development with Qwen & Ollama

Build AI Apps Locally



Riddhi Airan

• what is Qwen 2.5 ?

— LLM developed by Alibaba Cloud as part of the Tongyi Qiawen Series.

— Next-Generation AI model optimized to text understanding, generation, reasoning, and code assistance.

— Key features of Qwen 2.5

- Advanced Natural Language Processing (NLP)

↳ handles text generation, summarization, translation & Que/Ans.

- Multilingual Support

↳ works in multiple languages, including English & Chinese.

- Optimized for Code & Reasoning

↳ coding assistance, mathematical reasoning and logical problem solving.

- Efficient and Faster

↳ Improved response speed and lower memory footprint compared to its earlier version.

- Multiple Model Size

↳ used to balance efficiency and accuracy, enhance security & privacy.

- Enhanced Security & Privacy

↳ run locally on machines, making it Suitable for privacy sensitive applications.

• Qwen 2.5 Vs. Other Models

Feature	Qwen 2.5	Llama 3	GPT-4	Mistral
Developer	Alibaba Cloud	Meta	Open AI	Mistral AI
Language Support	Multilingual	English Focused	Multilingual	Multilingual
Fine-Tuning	Yes	Yes	limited	Yes
Efficiency	High	High	Moderate	High
Best Use Cases	Chatbots, Coding	Chatbots, NLP	Advanced AI	lightweight AI

Overview :- Qwen 2.5 is a great choice if you need a powerful, efficient and multilingual LLM running locally.

Riddhi Airan

• what is Ollama ?

— open-source runtime that enables easy deployment and execution of large language models locally on your machine.

— Simple interface to run, manage and interact with models like Qwen, Llama, Mistral, Gemma, and more.

— Key Features of Ollama :

- Local LLM Execution : Run AI models without cloud dependency.
- Model Management : Pull, run, and switch between different AI models effortlessly.
- Simple CLI Interface. Easy to use commands.
- Fast and optimized : Uses graph quantization for efficiency & faster response times.
- Developer friendly : Exposes the Rest APIs to integrate AI into Applications.
- Supports Multiple Models : Qwen, llama, Mistral, Phi-2, gemma and others.

Riddhi Airan

• Why Use Ollama ?

– Compare Ollama with Lang Chain and Open AI API.

Feature	Ollama	Lang chain	Open AI API
Runs Locally ?	Yes	No	No
Free to use ?	Yes	Yes	No (Paid)
Supports Custom Models ?	Yes	Yes	No (closed)
Optimized for speed ?	Yes	Varies	Yes

Overview :

Ollama is perfect if you want to run AI models locally, keep data private, and avoid API costs.

Riddhi Airan

• How Do Qwen 2.5 & Ollama work together?

- Step-by-step Process

1. Ollama manages and runs Qwen 2.5 locally.
2. User sends a prompt to Qwen 2.5 via Ollama's API
3. Qwen 2.5 processes the input and generates a response
4. The output is returned to the user through FastAPI or a chatbot UI.

Think of Ollama as the "engine" and Qwen 2.5 as the "brain" that powers your AI application!

Summary

- Qwen 2.5 + Ollama is a powerful combination for building AI applications locally without relying on cloud services like Open AI.
- Qwen 2.5 - A high-performance LLM for chatbots, coding, and automation.
- Ollama - A local AI runtime for running models securely and efficiently.

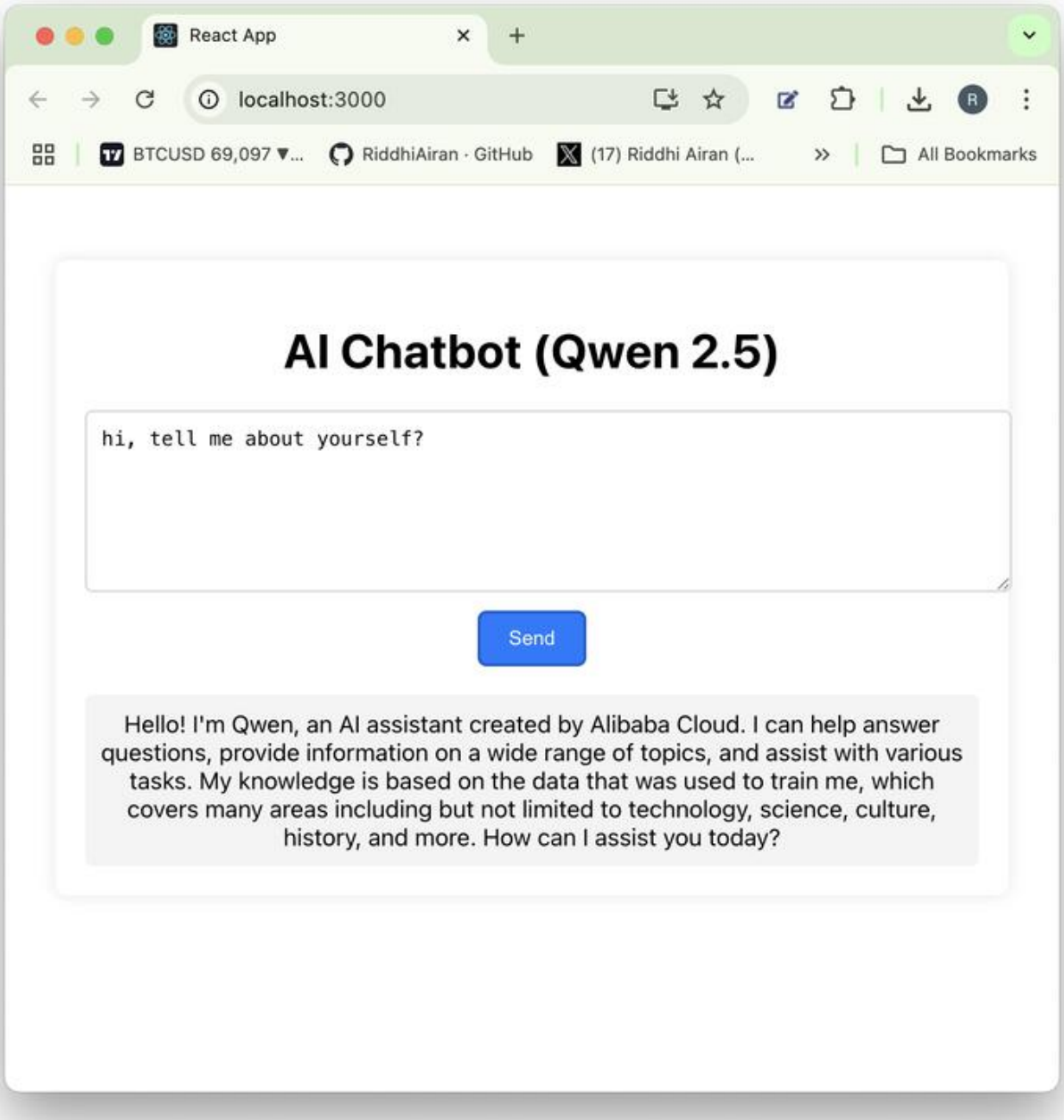
Riddhi Airan

AI Powered chatbot with Qwen 2.5 & Ollama

- Setting up Ollama and Qwen 2.5 on a local Machine.
  - install ollama
  - Download Qwen 2.5 Model
  - Test the Qwen 2.5 Model
- Using Fast API to create a Chatbot backend.
  - Set Up Python Environment
  - Create Fast API backend
  - Run the backend
- Creating a React.js frontend for user interaction.
  - Set Up React App
  - Create Chat UI
  - Style the Chat UI
  - Run the React App
- Deploying the Chatbot locally

Riddhi Airan

Deploying the Chatbot locally



Riddhi Airan