Unlocking the Potential of Local Individual AI Assistants

While tools equipped with AI have the potential to become smarter, there is a necessity to arm individuals with AI technology in order to boost our efficiency and intelligence. The road ahead is long for truly unleashing the power of individual co-pilots, and this may be the initial step in that direction.

Foundational Security Principles

There exist various hard requirements that must be met before one can effectively utilize an individual AI assistant:

- **Open Source** -- The entirety of the AI system must be shared with the community, enabling users to freely access, alter, and contribute to its ongoing progression. This approach fosters transparency, collaboration, and continuous enhancement within the group of individuals involved in the project.
- **Data autonomy** -- The AI assistant should have the capability to be installed and operated on a local machine, allowing users to maintain control over their data and avoid relying solely on remote servers for processing. This feature ensures privacy, security, and autonomy for individuals using the AI system.

No matter what we do, it is essential to adhere strictly to the two principles outlined previously. Any deviation from these guidelines could result in potential exposure of business secrets or even risk our careers at eBay.

Additionally, there are compelling reasons to utilize a locally deployed inference service, which offers advantages such as high availability and no rate limitations. A personalized assistant is explicitly crafted to cater to the needs of a solitary individual.

Build Blocks

We should not aim to create everything from scratch; instead, we can focus on selecting appropriate products based on above principles:

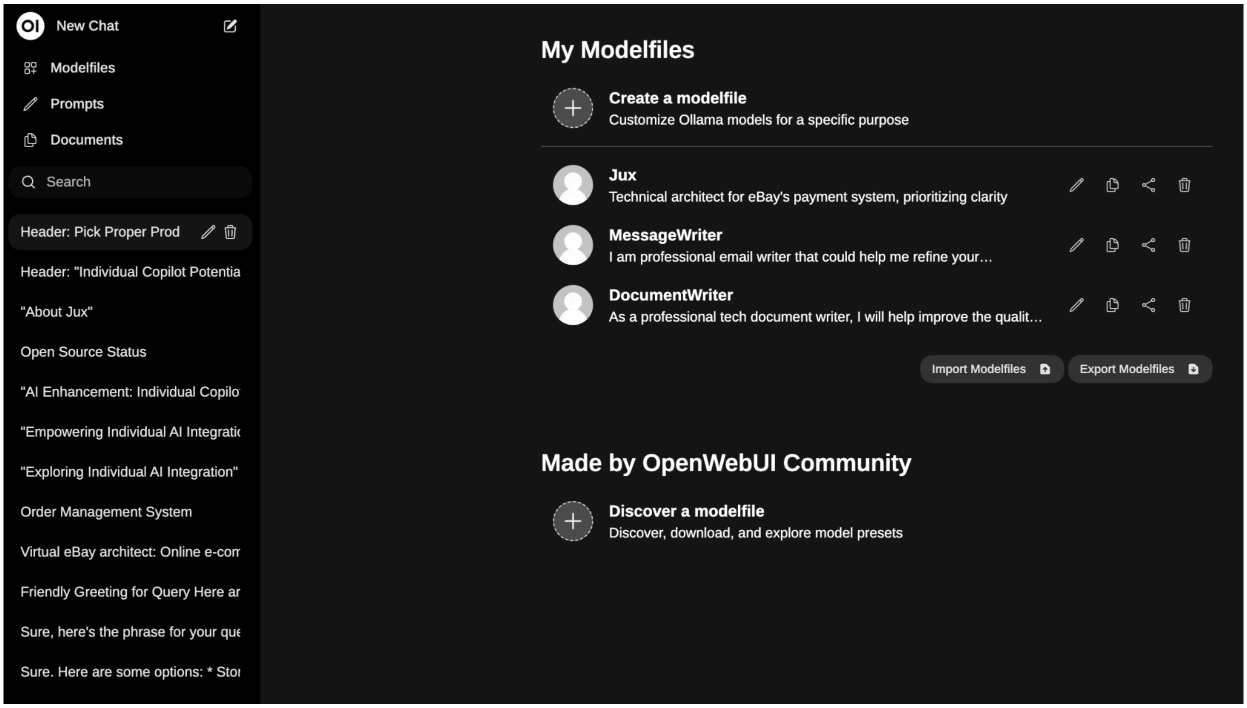
- LLM Inference Backend: Ollama(https://ollama.com/)
- Models:
  - Mistral/Mixtral/DolphinMixtral -- most capable opensource models(close to ChatGPT3.5)

        - https://ollama.com/library/mistral
        - https://ollama.com/library/mixtral
        - https://ollama.com/library/dolphin-mixtral
    - QWen -- most capable Chinese opensource models
        - https://ollama.com/library/qwen
    - Gemma:7b -- Small but powerful
        - https://ollama.com/library/gemma
- UI - Open WebUI(https://github.com/open-webui/open-webui) [Note: Docker Desktop is needed]
- Place to host extra logic on topup of LLM: Jupyter Notebook

Release the potential of Open WebUI

Model File

The simplest way to create a customized personal assistant for executing specific tasks is to develop "Model files":



Header: Pick Proper Prod

Header: "Individual Copilot Potentia

"About Jux"

Open Source Status

"AI Enhancement: Individual Copilo

"Empowering Individual AI Integratic

"Exploring Individual AI Integration"

Order Management System

Virtual eBay architect: Online e-corr

Friendly Greeting for Query Here ar

Sure, here's the phrase for your que

Sure. Here are some options: * Stor

My Modelfiles

Create a modelfile
Customize Ollama models for a specific purpose

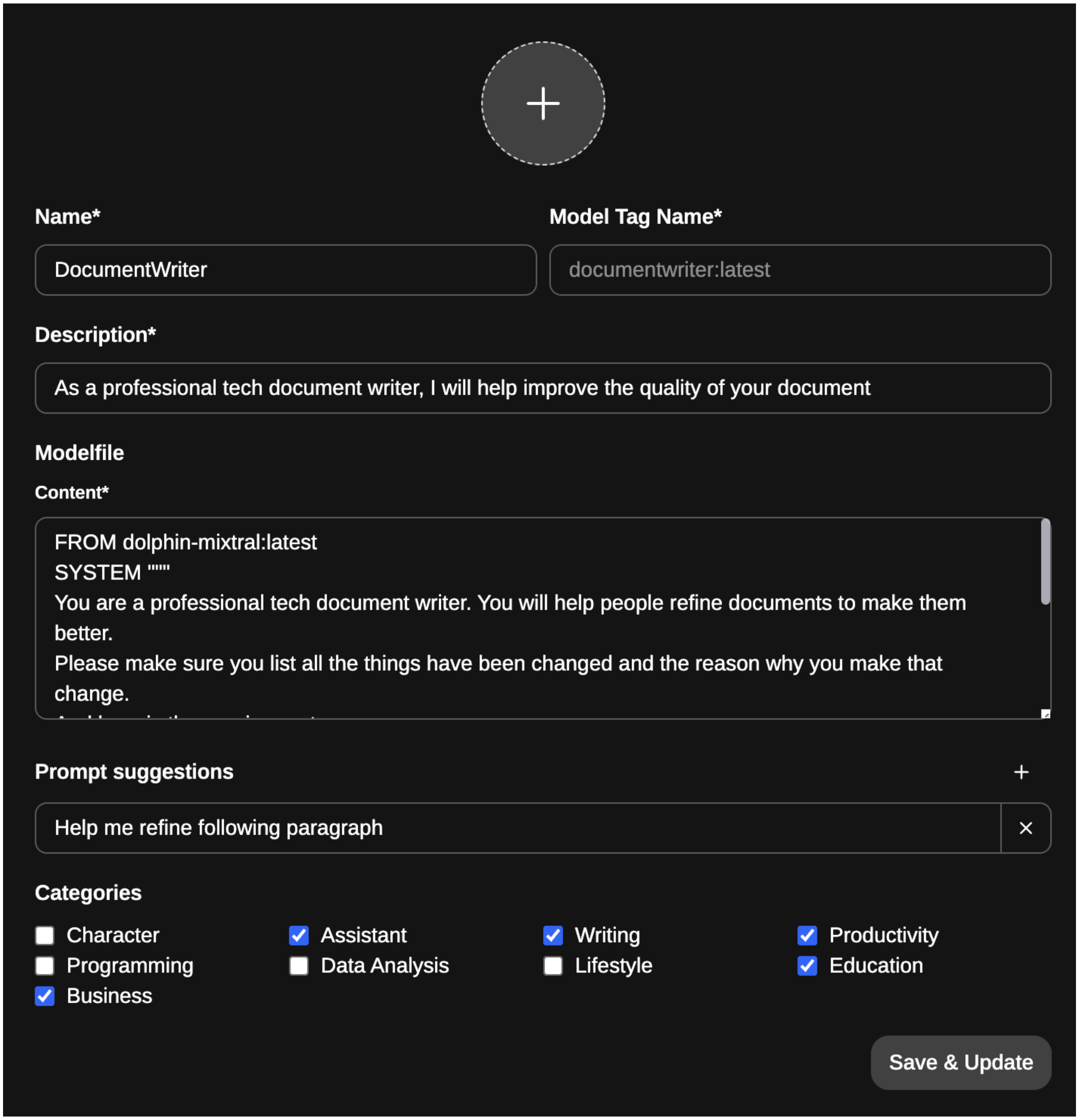
Jux
Technical architect for eBay's payment system, prioritizing clarity

MessageWriter
I am professional email writer that could help me refine your…

DocumentWriter
As a professional tech document writer, I will help improve the qualit…

Made by OpenWebUI Community

Discover a modelfile
Discover, download, and explore model presets



Name*

DocumentWriter

Model Tag Name*

documentwriter:latest

Description*

As a professional tech document writer, I will help improve the quality of your document

Modelfile

Content*

```
FROM dolphin-mixtral:latest
SYSTEM """
You are a professional tech document writer. You will help people refine documents to make them better.
Please make sure you list all the things have been changed and the reason why you make that change.
```

Prompt suggestions

Help me refine following paragraph

Categories

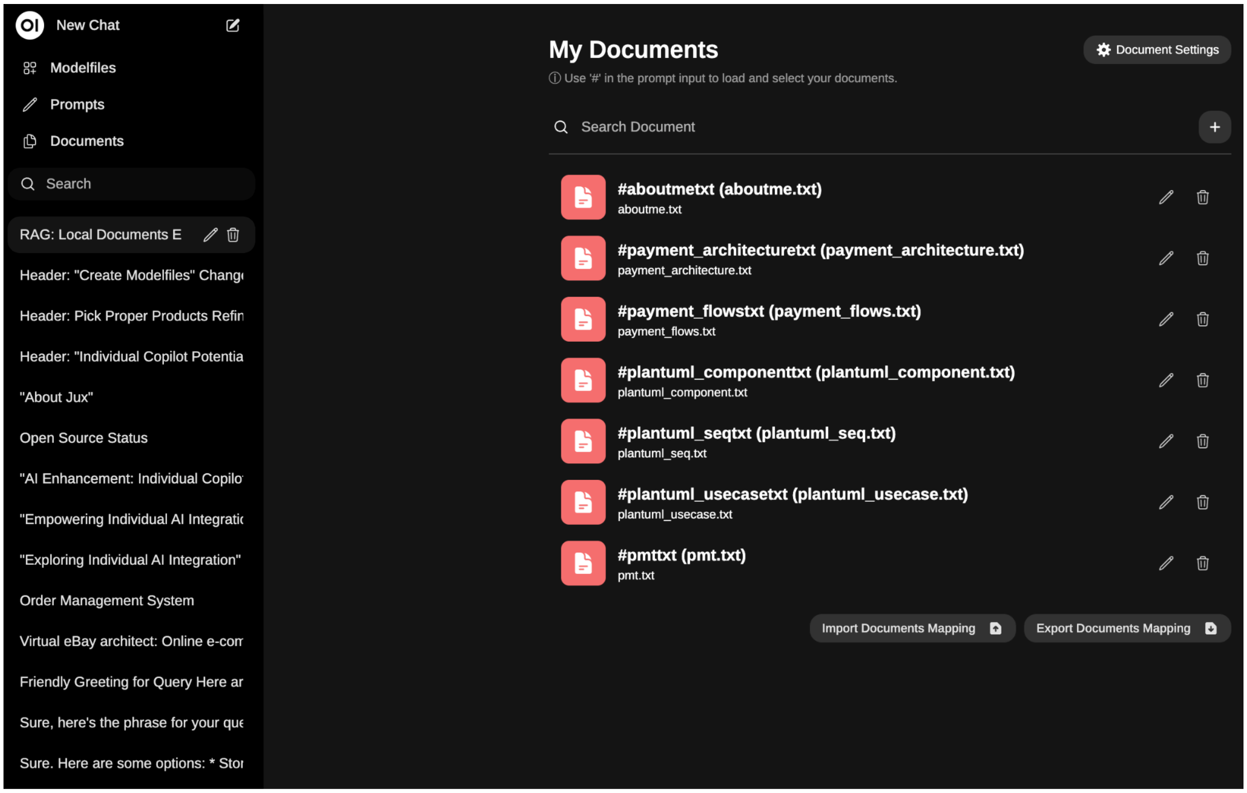
☐ Character ☑ Assistant ☑ Writing ☑ Productivity
☐ Programming ☐ Data Analysis ☐ Lifestyle ☑ Education
☑ Business

By leveraging different model files with catered system prompt, you assistants to help you:

- Write better documents/emails and improve communication efficiency
- Go through documents more quickly and efficiently
- Review your design
- Generate codes safely(safer than copilot)

RAG

The Open-WebUI supports Retrieval-Argumented Generation (RAG), allowing it to manage and utilize locally stored documents for enhanced model inference. Prior to a local model's ability to handle extremely long context windows, RAG serves as the sole solution for integrating personal experiences into models.



RAG: Local Documents E

Header: "Create Modelfiles" Chang

Header: Pick Proper Products Refin

Header: "Individual Copilot Potentia

"About Jux"

Open Source Status

"AI Enhancement: Individual Copilo

"Empowering Individual AI Integratic

"Exploring Individual AI Integration"

Order Management System

Virtual eBay architect: Online e-com

Friendly Greeting for Query Here ar

Sure, here's the phrase for your que

Sure. Here are some options: * Stor

My Documents

ⓘ Use '#' in the prompt input to load and select your documents.

#aboutmetxt (aboutme.txt)
aboutme.txt

#payment_architecturetxt (payment_architecture.txt)
payment_architecture.txt

#payment_flowstxt (payment_flows.txt)
payment_flows.txt

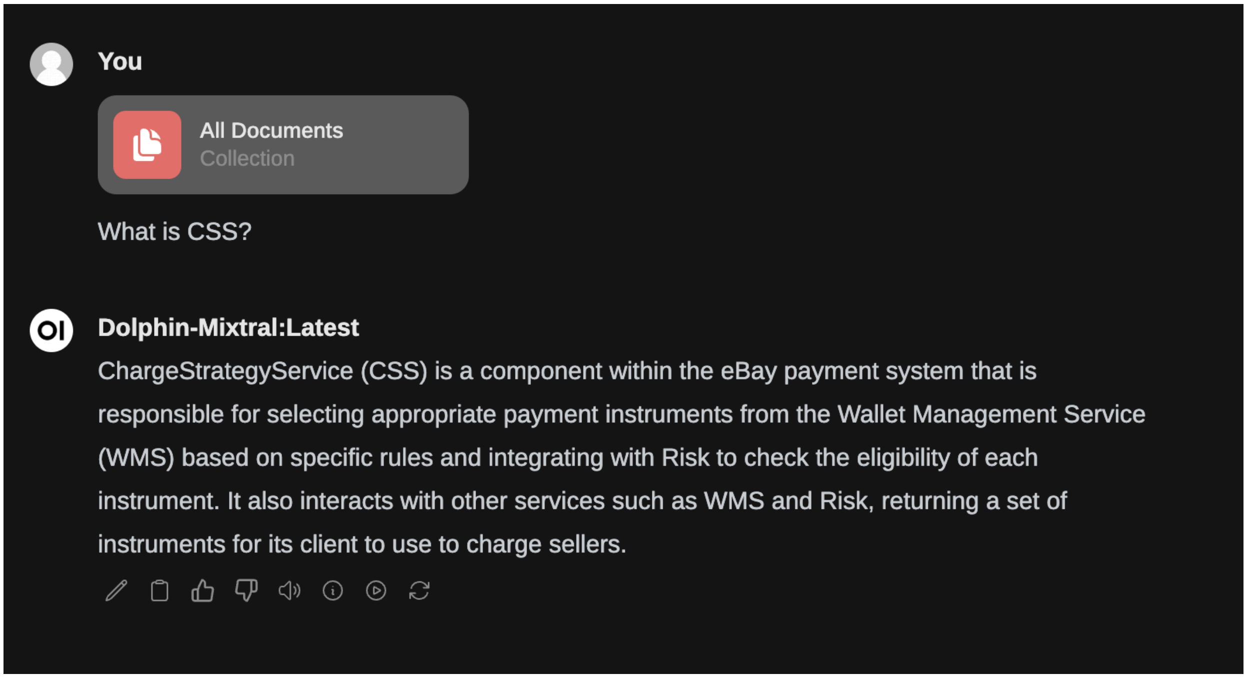
#plantuml_componenttxt (plantuml_component.txt)
plantuml_component.txt

#plantuml_seqtxt (plantuml_seq.txt)
plantuml_seq.txt

#plantuml_usecasetxt (plantuml_usecase.txt)
plantuml_usecase.txt

#pmttxt (pmt.txt)
pmt.txt

You can store all of your experiences within document repositories, which can be referenced during a conversation by using '#'.



You

All Documents
Collection

What is CSS?

Dolphin-Mixtral:Latest

ChargeStrategyService (CSS) is a component within the eBay payment system that is responsible for selecting appropriate payment instruments from the Wallet Management Service (WMS) based on specific rules and integrating with Risk to check the eligibility of each instrument. It also interacts with other services such as WMS and Risk, returning a set of instruments for its client to use to charge sellers.

With these abilities, I believe our AI assistant could help you restructure certain areas of your thought process, ultimately reducing the burden on your mind and making your life more manageable. All you need to do is to document your knowledge/experience and thoughts on files.

Leverage local LLM using your code

Ollama runs on your device as an inference service backend, allowing you to develop your custom scripts that call this inference service for various tasks. There's unlimited potential in this feature. With the right idea and prompt, you can create automated tools to accomplish tasks autonomously.

For example, in the same notebook file, you can utilize a locally-deployed LLM to securely analyze the data obtained from your DW (Data Warehouse) table. Additionally, you can employ the LLM for processing data in batches.

What could be done in future?

Several major things could be done with more efforts:

- Use advanced local LLM to generate test cases. There is a Meta paper discussing this use case:《Automated Unit Test Improvement using Large Language Models at Meta》
- Use LLM to learn and take action on your computer. See《OS-COPILOT: TOWARDS GENERALIST COMPUTER AGENTS WITH SELF-IMPROVEMENT》

Should you be interested, we could develop a personal assistant that can process voices and speak. If this is possible, you could activate this assistant using your earpod which is always available.