The Age of Data and Information Gluttony

Author: Wenceslao Alfageme
Published: 2 March 2025

Yes, I was (and still am) addicted to information and data consumption, whether on Artificial Intelligence, Data Management, Coding, Transformation, Economics, Finance, or other topics. Over the past few months, I have spent countless hours watching YouTube videos, reading thousands of articles (and some books too), keeping up with news pop-ups and reels, and diving into intensive research.

While this constant stream of data made me feel I was learning so much, I soon realised that my brain simply didn't have the capacity to retain it all and be able to produce an answer with enough specificity within the time frame I expected.

Developing a Smarter Approach to Data Retrieval

I was determined to build a system where I could retrieve facts and details on demand, so I needed to create a "**second brain**."

I explored multiple options available on the market. Many of these solutions were either too complex to integrate or came with high costs. This led me to experiment with an alternative approach: **leveraging local Large Language Models (LLMs) on my own laptop**.

I started with **Ollama** as my initial "exploration" platform, aiming to train these models using **Python** (programming language) to perform Retrieval Augmented Generation (RAG) on them. The results have been promising, I can now ask the LLM for specific information, and it provides detailed, accurate answers without hallucination (so far).

This approach seems particularly appealing due to its versatility, as my RAG interface is not confined to a single model. I could use local platforms like **DeepSeek** for tasks that require Chain of Thought, and others (like **Llama 3.1**, **Mistral**, etc.) for other simpler tasks, allowing for a broader application across different LLM environments.

I could also apply this approach when creating Multi-Agent Crews (e.g. **CrewAI**) to dynamically access different models based on their training data.

Additionally, by running these models locally, I can reduce the risk of data leaks, which is a major concern in today's data security landscape.

Let's Discuss: How Do You Tackle Information Overload?

I would love to hear your thoughts; what experiences or ideas do you have to further improve the challenges of information overload and data retrieval? Your insights could help refine this approach and contribute to a broader discussion on achieving smarter, more secure digital transformation. Please share your ideas in the comments below.

#DigitalTransformation #AI #Data #Information #LLM #ArtificalIntelligence #RAG #CrewAI

Definitions

A "**second brain**" refers to an external system, often digital, designed to capture, organise, and retrieve information much like our own brain, but with a greater capacity for storage and data processing. It helps manage and recall complex data, ideas, and insights, ultimately supporting decision-making and creative thinking.

Retrieval Augmented Generation (RAG) is a method that combines traditional retrieval of data with the generative capabilities of AI to produce more accurate responses.

"**Chain of Thought**" is the process of articulating a series of reasoning steps that lead from a question to an answer, helping to clarify and organise one's thinking.

CrewAI is a platform that enables the creation and deployment of AI agents, which automate tasks and streamline workflows by leveraging advanced machine learning and natural language processing.

In the context of this article, the following platforms and models represent a diverse ecosystem of tools that can be used to build a "second brain":

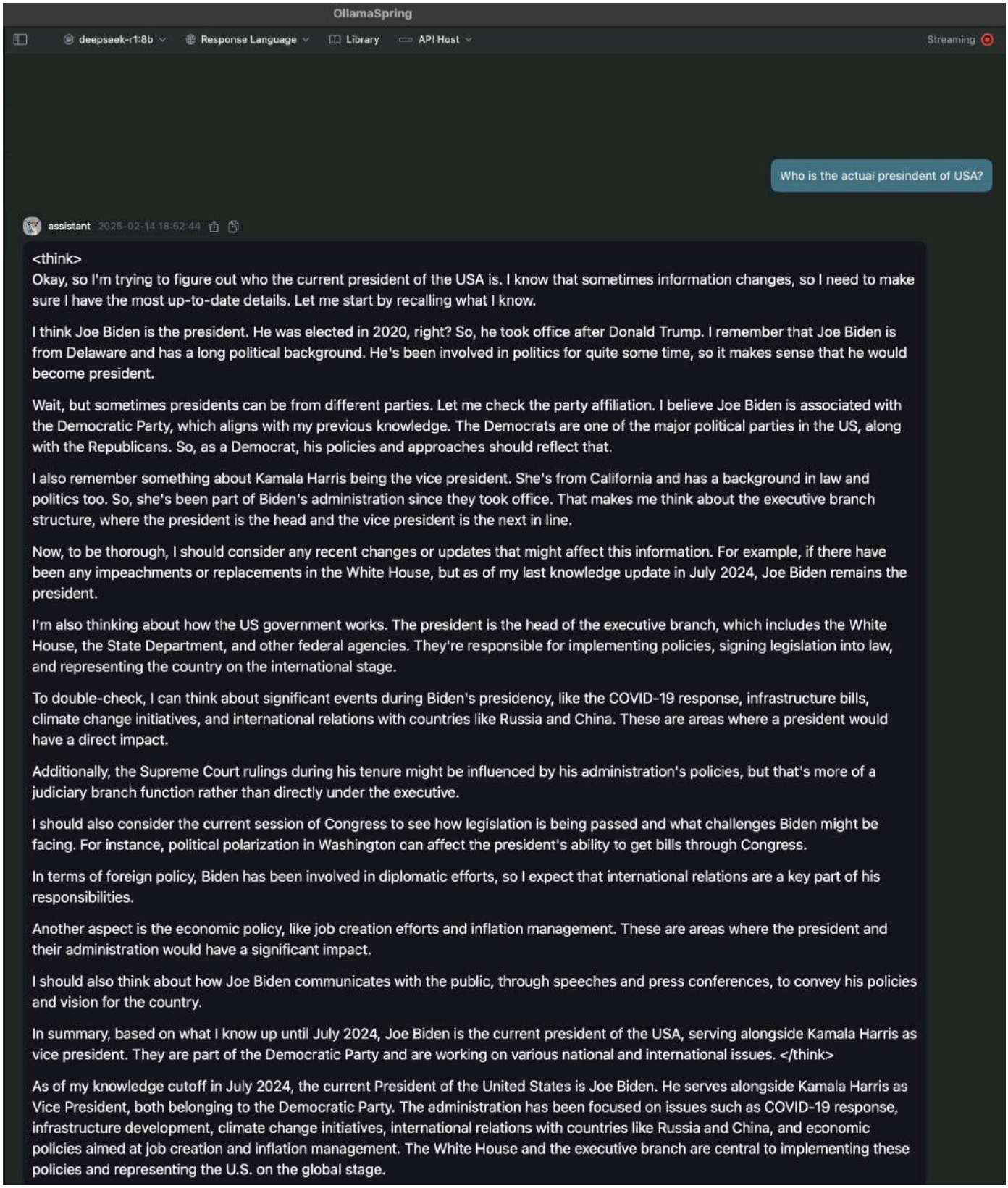
- **Ollama** is a platform that enables users to run local Large Language Models (LLMs) on their own machines, allowing for customised model training and usage without relying on external cloud services. This local setup enhances data security and customisation.
- **DeepSeek** is another platform that offers similar capabilities for retrieving and generating content using LLMs. It can also support Retrieval Augmented Generation (RAG) approaches, giving users the flexibility to work with different models based on their specific needs.
- **Llama 3.1** refers to a version of Meta's LLaMA (Large Language Model Meta AI) series, which are powerful open-source language models designed to perform a wide range of natural language processing tasks. In the context of the article, using Llama 3.1 locally means harnessing advanced AI capabilities for data retrieval and content generation, all while maintaining control over data privacy.
- **Mistral** is another state-of-the-art language model that competes with other leading models by offering high performance and efficiency. Like Llama 3.1, Mistral can be deployed locally, making it a viable option for secure and customised AI-driven solutions.

They offer various capabilities in terms of data retrieval, analysis, and content generation, all while providing the security and customisation required to mitigate risks such as data leaks.

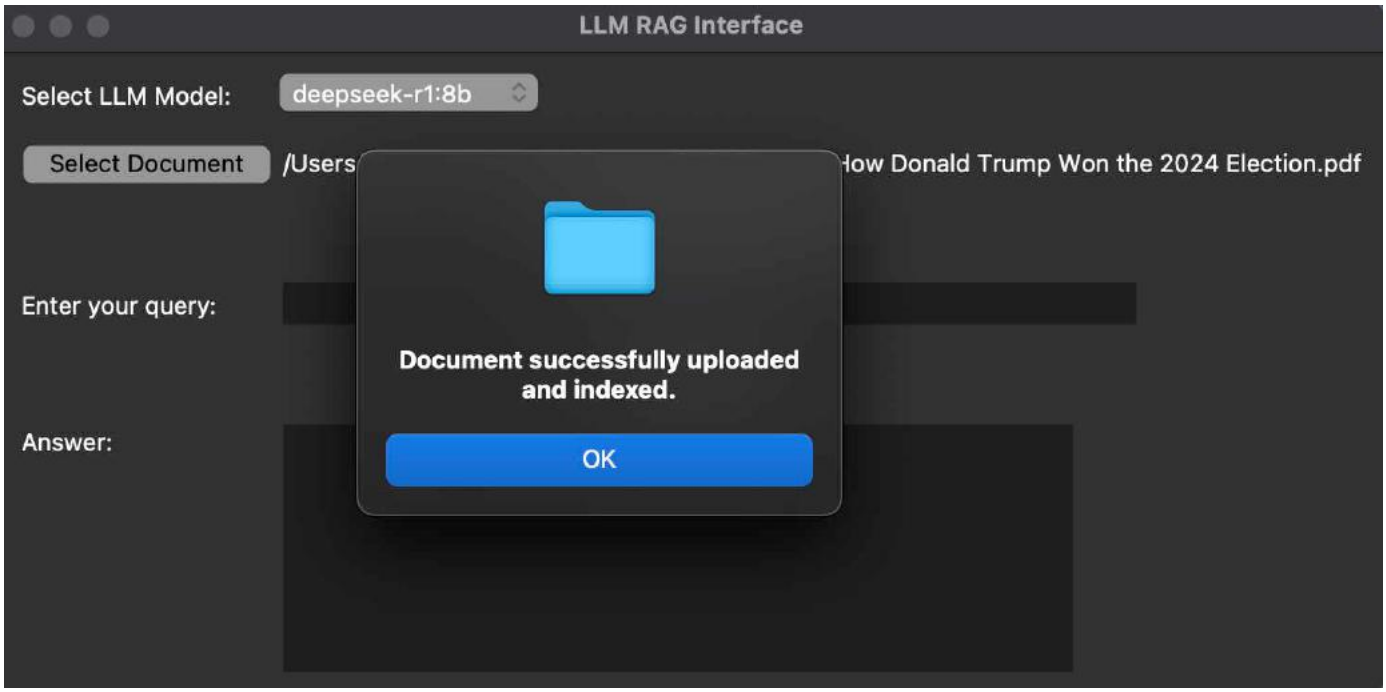
Experimentation Examples

See the screenshots bellow on how I ask my local LLM (trained with data up to mid 2024) about the current president of the US, and how it's response changes after the RAG was executed.

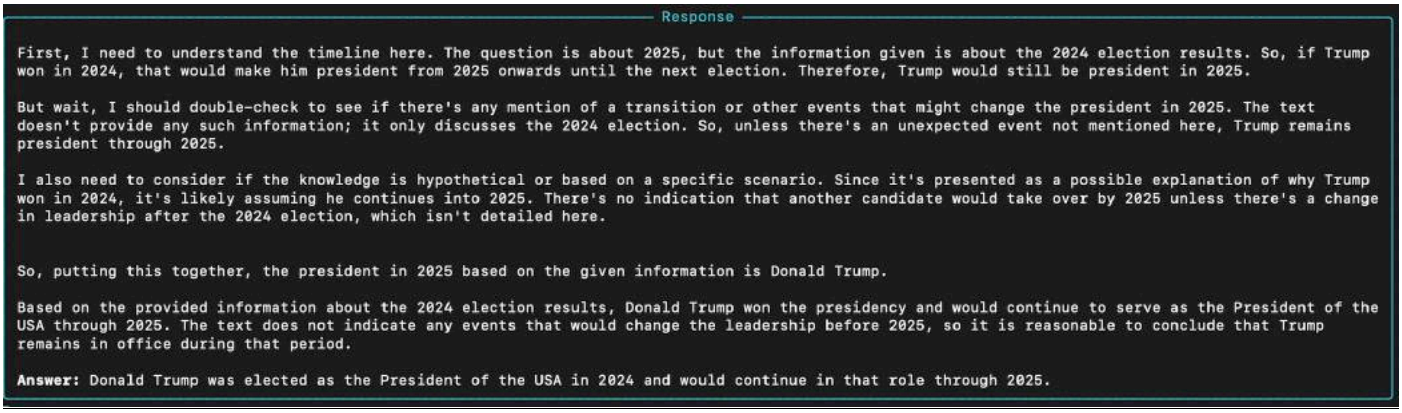
Step 1: Ask the DeepSeek LLM (pre-RAG):



Step 2: Use RAG to update the model with actual information



Step 3: Ask the DeepSeek LLM (post-RAG)



Disclaimer and Disclosure

Third-party Content and AI Assistance: This article references tools and software that are publicly available and proprietary to their respective creators. The author does not claim ownership or affiliation with these third-party products. This article was written by the author with assistance from Generative AI Language Models.

Transparency Notice: While every effort has been made to ensure accuracy, readers should verify information independently and consult official sources or documentation for the mentioned tools and software. The use of AI in the writing process is disclosed in the interest of transparency, but all opinions and analyses are the author's own unless otherwise stated.