Compound Artificial Intelligence Systems

Abhijeet Sahdev (as4673)
Submitted to: Dr. Guiling Wang

December 2, 2024

Contents

1 Abstract

This report explores Compound AI systems, where AI collaborates in a seamless, event-driven manner, tackling tasks with precision akin to nature's ecosystems. From Siri to ChatGPT, and Facebook to X, these systems form a life-like network, offering us a role as pivotal connectors to the universe. With a 10^{21} leap in computational power within 128 years [1], we're poised for a future where AI not only simplifies decision-making but also enriches this bond, all while ensuring we tread carefully and synchronously as we advance scientifically and engineer systems for social good.

2 Introduction

We examine the synergistic union of artificial intelligences, resembling a meticulously designed convergence of intellects, in an uninterrupted state when an event is triggered, dealing with intellectual tasks that match the system's caliber. Herein lies the potential for AI to first simplify decision-making, given the co-existence of data-driven facts without any substantial outcomes as in [2], and then possibly engaging in a nuanced understanding of complex phenomena, such as the enigmatic shifts in ocular pigmentation. Nonetheless, given the range of intelligent systems, from Alexa and Siri to Claude, Gemini, Grok, and social media platforms like Instagram and X, we observe an ecosystem of various agents, akin to a life-like ecosystem of nature and humans. Our role of acting as a common node on a graph of two sets, must instill a sense of paramount responsibility. This situation presents an unprecedented potential to drive insights using fundamental concepts of data engineering, machine learning, deep learning, and generative AI, enhancing our symbiotic relationship with nature through spiritualism and philosophical beliefs from various cultures on Earth, therein creating a profound relationship with our universe (Fig1). In a way, the interplay between the Golden Ratio, the Fibonacci sequence, and the Yin-Yang can be seen as a cosmic dance of numbers and philosophy, where each element complements the other, creating a universe that's not just mathematically beautiful but also philosophically profound. With optimism, we now look towards a promising and secure future by being mindful of the equal or sequential footing of catastrophic events often coinciding with scientific advances. Keeping these precautions in mind, we now delve into recent developments in bridging this gap between the universe and our understanding of its various elements through compressed versions of the entity, Internet, using a compound AI system.

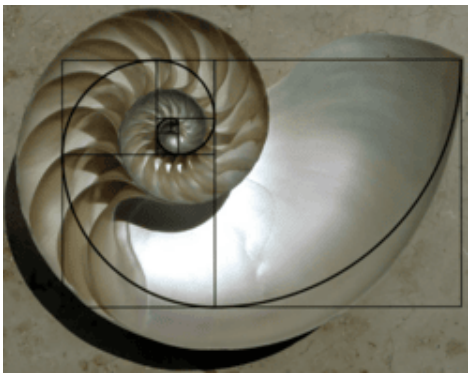
3 Present Day State

The World Wide Web as an ecosystem or a food chain has its energy flowing from one entity to another through data. As we may know, data is primarily of two types:

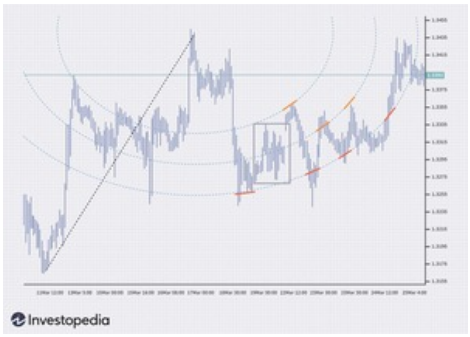
- Structured Data, example spreadsheets with rows and columns. Consider a credit card approval system which is an automated task through various banking applications. A new student in a foreign country would find it nearly impossible to receive a credit card, unless there's a banking institute in their home country, establishing a link for a credit history to be transferred. At times, from personal experience, in-spite of establishing a clear audit using debit cards, certain banks decline such approvals based on the available bank balance or their inability to verify a user's identity, given the latent space of the user's interactions in their specific country or jurisdiction.

- Unstructured Data, example a bag of nodes could represent a sense of cohesion using similarities at times using attributes such as location, emails, legal names and surnames on government issued

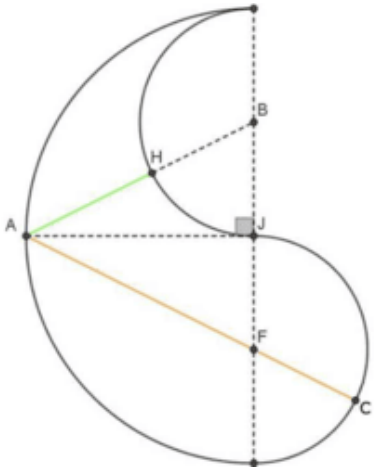
(a) Spiral Galaxy by NASA, [3]



(b) Nautilus shell by Chris 73 / Wikimedia Commons, [3]



(c) Fibonacci Arcs[4]



(d) Taoism [5]

Figure 1: The Cosmic Dance of Mathematics and Philosophy

Figure 2: Haystack Overview [6]

cards. Clearly, a global database could easily track and monitor every user and could use the same automated system to approve simple credit card requests, track money laundering, etc.

As drawn out using through first thoughts, the banking industry alone has various endpoints, each requiring a different context base to carry out the desired tasks. The backlogs are alarming when it comes to approving visas for over-qualified students who chase the American Dream or living in another country, which could easily be automated using a student's achievements, papers published, degrees awarded and miscellaneous achievements through recognized and verified institution. A compound AI system, like a simple Gradient Descent Algorithm, depends on the quality of the data it is trained out. Thereby, scoping or taking in the desired context offers a streamlined approach, introducing Retrieval Augmented Generation (RAG) applications. We have multiple chat-bots that have been trained on unfathomable data if precision is considered to be the most imperative guideline. Social Media applications offer analytics for user interactions and clearly quantifying the revenue that users can generate using it but its energy consumption is difficult to quantify even today by available open source software discussed in Section 3.2 as it still faces challenges in accurately monitoring cloud systems.

3.1 Haystack

It is an open-source end-to-end framework, similar to React/React-Native by Meta and Flutter by Google when coupled with cloud consoles, utilized for building production-ready Large Language Model (LLM) Applications with RAG pipelines and state of the art search systems that work over large documents. Given that they utilize LLMs for their tasks, it is understood that they work intelligently, to the best of their standards for ensuring efficiency in terms of both, cost and performance. Fig 2 depicts that it has been designed in a modular way, allowing the users the flexibility to use any LLM of their preference. A sample Haystack pipeline is depicted in Fig 3.

3.1.1 Abstract Units

Components: Haystack offers a range of components, each designed for specific functionalities such as text generation or document retrieval, facilitating to design the system in an agile fashion. They are present as python classes.
Generators: These components are responsible for text creation, interpreting user prompts to generate responses. They are categorized into chat-oriented generators for conversational engagements and non-chat generators for tasks requiring direct text output, like translation.

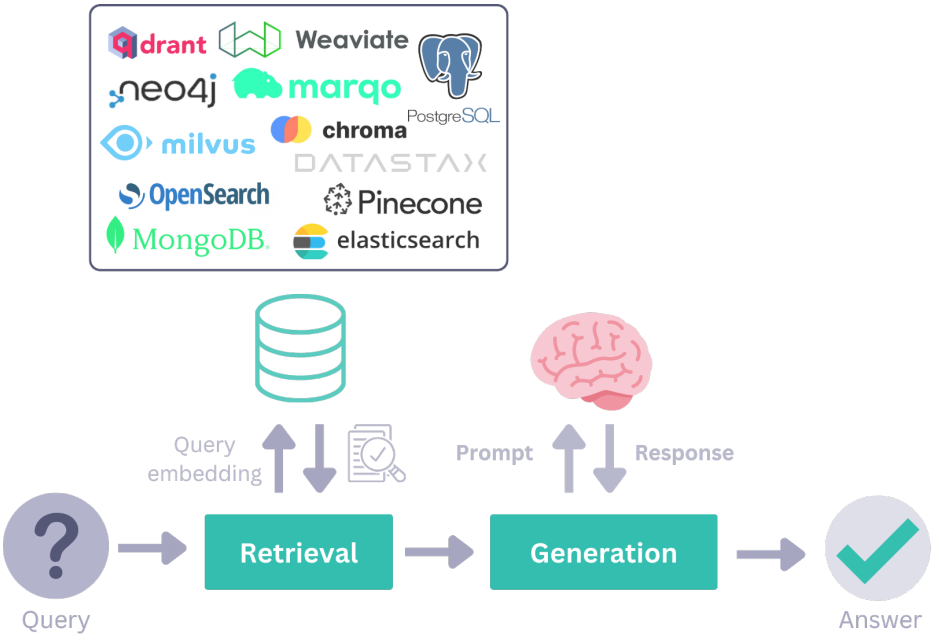


Figure 3: Pipeline [7]

Retrievers: These are document retrieval agents, similar to APIs in web applications but with an added search functionality through extensive databases to locate documents pertinent to the user's inquiry, thereby ensuring precise responses.
Document Stores: These serve as the primary repositories for documents, accessible by various components for both reading and writing operations, ensuring the system's data organization, similar to modern day data warehouses.
Data Classes: These are integral for data transfer within Haystack, managing the flow of information such as text, metadata, or answers across different parts of the system.
Pipelines: Within Haystack, pipelines orchestrate the workflow by integrating various components into sequences for tasks like data processing or querying, allowing for storing and sharing user defined processes.

3.2 Code Carbon

A common rhetorical statement raised today is the ecological impact of using complex systems that utilize enormous amounts of computing power, here comes CodeCarbon [8], another open source tool designed to measure the carbon footprint of machine learning / deep learning experiments, addressing this concern. The software uses global carbon intensity data from sources like ElectricityMaps, the European Environment Agency (EEA) or OWID [9] to raise the awareness of developers, researchers and companies about their ecological impact of their work.

3.2.1 Carbon Footprint Calculation

Carbon dioxide emissions (CO_2eq) are calculated as $C \times E$ where :

- C represents Carbon Intensity of the electricity consumed. Units: g of CO_2 emitted per kilowatt-hour of electricity.
- Energy Consumed by the computational infrastructure. Units : kilowatt-hours.

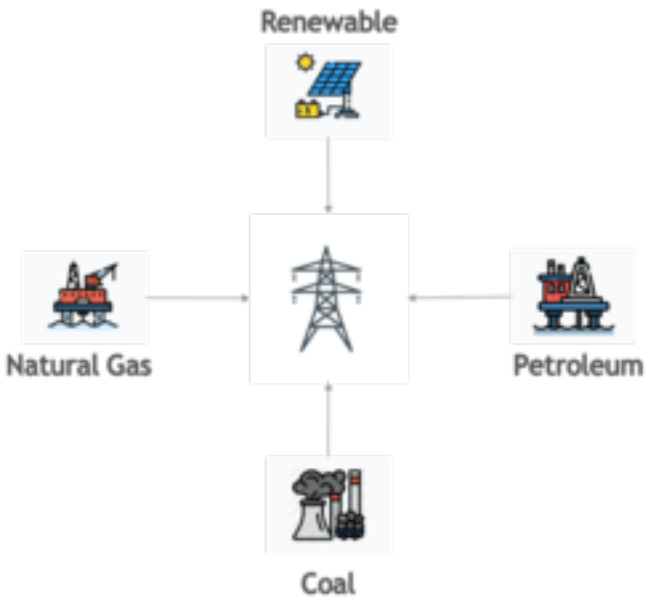


Figure 4: Carbon Intensity [10]

Definitions and examples can be viewed on [10].
Energy Consumption refers to the energy consumed by hardware during the execution of code. This includes CPUs, GPUs, and other components, and caters to different companies including NVIDIA, INTEL, Windows or Mac(Intel) as well as Apple Silicon Chips (M1 and M2). It uses a scheduler to log the energy consumed in intervals of 15 seconds. Whereas, carbon intensity of consumed electricity (Fig4) is calculated as a weighted average of emissions used to generate electricity from various sources including non-renewable and renewable energy sources, which varies by location as stated earlier.

4 Future of Scope of Work

Given these many options and various ways to organize them, it is imperative for us design one system, which is a software engineering task using agile software development paradigms, involving various inventors and teams across the globe will be a promising path forward. While energy consumption is one factor that directly affects us, the users or in scientific terms, an agent interacting with the system, value their privacy now more than ever. For example, Grok generates his output using my recent tweet and communicates in a way that is hilarious to me. This invasion of privacy as others would call it is fine with me as I practice and embody absolute transparency in my time on the World Wide Web, obviously not in terms of passwords. Nonetheless, not everyone is an outlier and my day to day web-surfing brought Sir Tim Berners-Lee vision of empowering users to be in-control of their data, but more importantly their digital traces of their interactions on the Web by introducing a concept called PODS [11], which essentially is a safe-lock for a user's data. Personally, I see this as an absolute win-win in terms of privacy and storage space, creating another progressive step forward by freeing up space to store intelligent systems and its computational tools on remote servers. However, this calls for a total revamp of the World Wide Web which has its downsides as well. A step towards a soft-spot in resource utilization could be to "enable" a link between our personal computers, in terms of its computing power and services such as Colab [12] offered by Google, similar to Amazon's Web Services console [13].

References

[1] Steve Jurvetson. *The Moore's Law Update*. X post. Dec. 2024. URL: https://x.com/FutureJurvetson/status/1863649174358831312.

[2] Bernie Sanders. *The Pentagon's "mishap"*. Dec. 2024. URL: https://x.com/sensanders/status/1863268770371772863.

[3] CNET. *Nature's Patterns: Golden Spirals and Branching Fractals*. Pictures. URL: https://www.cnet.com/pictures/natures-patterns-golden-spirals-and-branching-fractals/.

[4] Investopedia Staff. *Elliott Wave Theory: Theory and Practice*. Mar. 2004. URL: https://www.investopedia.com/articles/technical/04/033104.asp.

[5] Archimedes-Lab. *Golden Ratio and its Inverse in Yin-Yang*. Oct. 2019. URL: https://archimedes-lab.org/2019/10/06/golden-ratio-and-its-inverse-in-yin-yang/.

[6] Deepset. *Haystack - Question Answering over Documents and Databases*. https://haystack.deepset.ai/overview/intro. Accessed: [August 21, 2024].

[7] Deepset. *Haystack Documentation - Components Overview*. Accessed: 2024-12-02. 2024. URL: https://docs.haystack.deepset.ai/docs/components_overview.

[8] CodeCarbon Contributors. *CodeCarbon*. 2024. URL: https://mlco2.github.io/codecarbon/index.html.

[9] Energy Institute - Statistical Review of World Energy (2024) Ember (2024). *Our World in Data, Carbon Intensity of Electricity*. 2024. URL: https://ourworldindata.org/grapher/carbon-intensity-electricity#explore-the-data.

[10] CodeCarbon Contributors. *Methodology*. 2024. URL: https://mlco2.github.io/codecarbon/methodology.html.

[11] Tim Berners-Lee. *The Web Can Be a Force for Good. Here's How*. The New York Times. Jan. 2021. URL: https://www.nytimes.com/2021/01/10/technology/tim-berners-lee-privacy-internet.html.

[12] Google. *Google Colab - Free Jupyter Notebook Environment*. Accessed: [Dec 3, 2024]. 2024. URL: https://colab.research.google.com.

[13] Amazon Web Services. *Amazon Web Services (AWS) - Cloud Computing Services*. 2024. URL: https://aws.amazon.com.