The Unstoppable March of Artificial Intelligence: The Dawn of Large Language Models

Umair Ali Khan

Ever since the significant leap in Artificial Intelligence (AI), triggered by deep learning around 2012, the field has been on a relentless march forward. The watershed moment of deep learning sparked a revolution in AI. Since then, there's been no stopping the tide of innovation. AI has been growing at a breakneck pace, with developments pouring in at an unprecedented rate.

One of the most notable advancements within this AI explosion has been the birth and rise of Foundation Models (FMs) (Vastola 2023). FMs are trained on broad data that can be adapted to a wide range of downstream tasks, and can handle a multitude of data and modalities (Naz 2023). However, these models are intended to serve as a basis for (foundation) and not to be used for a particular end goal. Large Language Models (LLMs), such as chatGPT, are a type of FMs trained specifically on language-related data (text). Fed with a large amount of data sources including books, articles, scripts, and the vast expanse of the web, LLMs can understand the subtleties between words, phrases, and sentences. They learn to decipher the patterns, the grammar, and the semantics that form our language, resulting in their ability to generate responses that are not just coherent and contextually relevant, but also human-like. The distinction between FMs and LLMs is rather tacit, and the terms are used interchangeably.

The emergence of LLMs has revolutionized human-machine interactions. Since the introduction of transformer-based models like BERT and GPT in the mid-2010s, and notably, ChatGPT in 2022, machines have evolved from functional tools to conversational partners. The advent of the 'attention' mechanism in 2017 (Vaswani et al. 2017) marked a significant milestone in the evolution of transformer models, substantially enhancing their performance. This transition broadened the application of LLMs beyond the realm of research, introducing the public to AI models capable of meaningful dialogues, query resolution, and creative text generation. The resulting shift in our AI perception has catalyzed numerous opportunities for businesses, academia, and individual users, effectively integrating AI into our daily lives.

The Blossoming AI Landscape

AI applications using LLMs have exploded since chatGPT's debut and are no more limited to conversational agents. Talking specifically to education and research, the future of academia is likely to be transformed by AI language models. LLMs are being used in a plethora of education and research applications. For example, typeset.io can summarize a research paper in simple language, explain any selected part of the research paper including mathematical expressions, and search for required information from a paper. Elicit.org is an AI research assistant that is ideal for evidence synthesis and text extraction.

Elicit pulls publications from Semantic Scholar and expedites the literature review process. Users enter a research question into the search box and the AI attempts to identify the top papers in the field, summarize takeaways from the paper, and extract key information into a research matrix. Elicit can also fetch the citations for a given paper with a specific discussion, highlighting the critical comments on the paper.

Scite.ai is a platform that offers a range of tools for academic research, including a chatbot that provides access to a database of over 1.2 billion scientific articles. The tool contextualizes citations and distinguishes if the citation was made in the introduction, methods, results, or discussion section. Consensus is a search engine that uses AI to extract and distill data straight from scientific research in order to provide users with evidence-based answers to their queries. Consensus only searches through peer-reviewed, published sources and can help users save time and energy by providing them with accurate and condensed summaries of studies.

Scinapse is another AI-powered academic search engine designed for research purposes. It allows users to search for research papers in major STEM journals, providing a preview of the author explorer, which allows users to browse authors by research field, affiliation, or country. Research Rabbit is a powerful AI research assistant that finds and organizes research papers for you and your collaborators. Enter a keyword or phrase, and Research Rabbit will return a list of relevant papers through Semantic Scholar or PubMed search. ReadCube Papers is an AI-powered reference manager and citation software that helps you access scholarly articles from anywhere. The tool makes it easy to read, annotate, and share articles on any device, so you can work on your research wherever you are.

DiaChat is a web application that allows users to create and edit diagrams using natural language processing. Users can simply describe what they want their diagram to show, and the application will generate the diagram for them, eliminating the need for manual placement and layout. Education Copilot is an AI-driven tool designed to streamline lesson planning and material creation for educators. It offers a variety of AI-generated templates for lesson plans, writing prompts, educational handouts, student reports, project outlines, and more.

These are just a few examples of AI-powered tools for education and research. Dozens of new tools are being introduced on a daily basis in a variety of fields. Some platforms which provide a curated list of such AI tools include futuretools.io, aitools.fyi, topai.tools, aitoolsdirectory.com, and futurepedia.io, to name a few. A curated list of tools and resources regarding the GPT-4 language model can be accessed at gpt4.tools. A very comprehensive curated list of resources dedicated to open-source GitHub repositories related to ChatGPT can be found at this link. A list of AI resources (courses, tools, apps, open-source projects) can be accessed at this link. Another curated list of open-source AI tools and resources can be accessed at this link. These resources can streamline the research process by offering quick access to a wide array of AI tools and platforms.

Multimodal LLMs (MLLMs) are also making strides in the AI landscape. MLLMs can understand not only text but also various other forms of data, such as images, sound, and more (Zapier 2023). These models convert multimodal data into a common encoding space, which means they can process all types of data using the same mechanism. This allows the models to generate responses incorporating information from multiple modalities, leading to more accurate and contextual outputs. A curated list of MLLMs and other related resources can be accessed at this link. These open-source MLLMs can be used

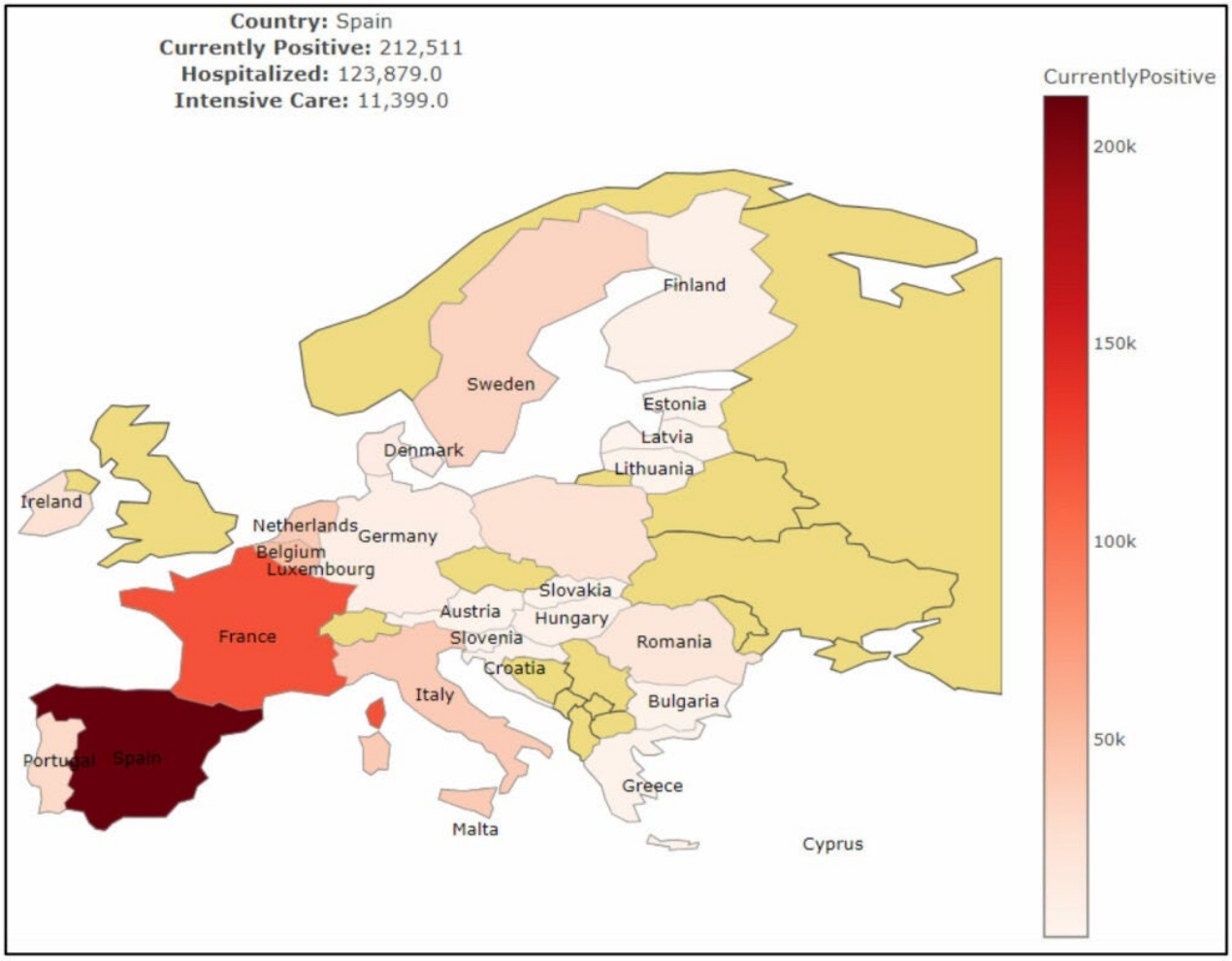
for developing a number of applications such as domain-specific conversational agents, match-making platforms, translation services, emotion/sentiment analysis, content generation, personalized healthcare, improving learning experience in education, document analysis, and personalized advertising, to name a few. Recently, an open-source unified framework for multimodal learning has been developed which can handle a wide range of tasks including fundamental perception (text, image, point cloud, audio, video), practical application (X-Ray, infrared, hyperspectral, and IMU), and data mining (graph, tabular, and time-series) (Zhang et al. 2023). Such models can have huge potential in healthcare for early disease detection, analyze hyperspectral images for precision agriculture, improving object detection and navigation capabilities in autonomous vehicles using point cloud data, predicting market trends in finance, processing graph data in social platforms to reveal intriguing user behavior insights, and content moderation across digital platforms, to name a few. Models tailored to specific domains, offering concentrated, specialized, and precise insights about a particular field, are also gaining traction. For example, a Multimodal model mPlug-owl has been recently used for OCR-free document understanding (Ye et al., 2023).

On July 6th, 2023, OpenAI further enhanced the capabilities of chat GPT by introducing a 'Code Interpreter' plugin. This feature acts as a digital data analyst, readily accessible to perform a range of tasks such as data cleaning, analysis, in-depth explanations and insights, visualization, video editing tasks like trimming or format conversion or even solving mathematical problems. Furthermore, the Code Interpreter enables direct uploading of local files and datasets to ChatGPT, accommodating various formats. This upgrade essentially brings the power of data analysis to your fingertips, greatly enhancing the user's ability to interact with and manipulate data efficiently and effectively (for Security and (CSET) 2023)..

To check the capabilities of Code Interpreter, I obtained the most recent Covid-19 dataset from the Humanitarian Data Exchange website, featuring 309,902 day-by-day records from 224 countries. The dataset comprises various parameters including the count of positive cases, fatalities, recoveries, current patients, hospitalizations, intensive care unit (ICU) admissions, along with geographic coordinates, date, and country.

I fed this dataset into the Code Interpreter, first prompting it to 'Use the file I uploaded and do the basic analysis'. Despite the absence of metadata, it began by providing an in-depth analysis of the dataset including a comprehensive description of all variables, their data types, missing data, and other statistics. Subsequently, I prompted it to 'Give me 10 ideas for trends, visualization, and analysis I can perform with this data'. After it gave 10 analysis ideas, I prompted it to 'Do all these. ', and Viola! It performed all the analyses with advanced visualizations complemented by thorough explanations and insights. I am only sharing the visualizations in this PDF-file.

We can ask it for several other analyses by writing the relevant prompts. For example, I wrote the following prompt: 'Create an interactive European map to show the data of each European country on 1st June 2020 with the following statistics: Currently Positive, Hospitalized, and Intensive Care. The interactive map should be plotted using Plotly or some other suitable library. The map should be able to zoom in/out and show the details at mouse hovering.' It created the following interactive map.



The Advent of AI-Based Search Engines

Although, Google and other search engines embraced AI to optimize ranking and a better understanding of the context of words in search queries, a major drawback of these search engines is that they produce generic, non-personalized results which require users to sift through a vast array of web pages to find the information they need. The concept of using LLMs for search engines is becoming increasingly popular. Several search engines such as Microsoft Bing Chat, Perplexity AI, Komo, Waldo, Andi, and You have integrated LLMs into their search capabilities. OpenAI too has also introduced the 'browse with Bing' option in ChatGPT.

AI search engines can understand the intent and context behind a search query, and can provide more relevant, precise, concise, and personalized results based on user behavior and preferences with citations. AI-powered search engines can also use data from multiple sources to provide richer and more accurate search results, and can learn and adapt to user behavior over time to further improve the quality of search results.

So what is the difference between conversational agents like ChatGPT and AI-powered search engines? ChatGPT excels at generating coherent, human-like responses to user prompts. Its ability to mimic natural conversation can make it more engaging and user-

friendly. However, its responses are limited by the data it was trained on, and it can "hallucinate" plausible but incorrect answers. Additionally, integrating up-to-date information is a challenge, as retraining the model with new data can be expensive.

Currently, the most prevalent AI-powered search engines are Microsoft Bing Chat, Perplexity AI, and chatGPT with browsing support (At the time of writing this article, ChatGPT with browing support is temporarily or permanently disabled). Each of them has its pros and cons.

While Perplexity doesn't restrict what browser you use, Bing Chat does, forcing you to use Microsoft Edge. In terms of technology, Bing Chat uses GPT-4, while Perplexity AI uses GPT-3 and GPT-4.

Perplexity AI with default GPT-3 model has the fastest response and decent response quality. However, with the GPT-4.0 model (also known as copilot) with more context awareness, its response time is comparable to Bing Chat. On the other hand, Bing's responses come with follow-up questions which give an enhanced conversational experience. While ChatGPT with browsing support may take more time and occasionally struggle to find answers to queries, when it does succeed, it offers greater depth and clarity in its responses. On the other hand, Bing Chat asks for starting a new topic after each 6 questions. Also, Bing Chat stops responses abruptly if it finds a question inappropriate. While Perplexity AI is serviceable as a research tool and is highly accurate, Bing Chat is better at holding conversations and sounds more natural (Ibrahim 2023). That being said, it is well worth trying out all these search engines to discover which one best suits your needs and preferences (Naz 2023; Vastola 2023).

Choosing the Right Tool: Conversational Agent vs. AI Search Engine

The choice between using a conversational agent like ChatGPT or an AI search engine depends on your needs. ChatGPT is ideal when you want a more interactive and conversational experience, or when you need to generate diverse forms of text, from prose and poetry to computer code. It's also useful when you need a tool that can understand and respond to prompts in a human-like manner.

On the other hand, an AI-powered search engine might be more suitable when you need to find specific, up-to-date information from the internet, especially when it has a wide context that is beyond the capabilities of traditional search engines. These search engines are also more effective at handling large-scale information retrieval and organization tasks.

For example, 'to create a table of 2022 data of the top 10 happiest countries along with their happiness index, life expectancy, health expenditure per capita, and the average number of work hours per week', I would refer to an AI search engine. This information cannot be fetched by the Google search engine in a single go. We cannot also use ChatGPT conversational tool since its knowledge cut-off date is September 2021. Hence, We would either use Bing Chat, Perplexity, or ChatGPT with Bing. For the mentioned query, the copilot mode of the Perplexity AI creates the following table with references.

Country	Happiness Index	Life Expectancy	Health Expenditure per Capita (USD)	Average Work
Finland	7.8	82.1 years	5,267	40
Denmark	7.6	81.7 years	5,703	37
Iceland	7.6	83.1 years	5,292	40
Switzerland	7.5	83.6 years	9,047	41
Neatherlands	7.5	81.9 years	6,052	30
New Zealand	7.4	82.8 years	4,882	37
Israel	7.3	83.3 years	2,998	43
Norway	7.3	83.3 years	7,347	37
Sweden	7.4	82.9 years	5,715	36
Luxembourg	7.2	82.4 years	8,684	40

Sources:
Happiness Index:The World Happiness Report
Life Expectancy:World Population Review
Health Expenditure Per Capita:World Bank Data
Average Work Hours per Week:Statista

Note that some of the data may be from different years, as the most recent data available varies by source

Here is a table of the top 10 happiest countries in 2022, along with their happiness index, life expectancy, health expenditure per capita, and average number of work hours per week:

AI-powered search engines can be utilized in several tasks. For instance, finding relevant academic references through AI-enabled search engines is immensely helpful for the literature review. Consider the relevant academic references fetched by Perplexity AI's copilot for the query: 'Find the relevant academic references for 'interactive cinema' or 'emotional cinema'.' These screenshots show the responses of Perplexity ai and Microsoft Bing Chat. While both Perplexity AI and Bing Chat delivered 6 references, the ones from Perplexity AI were notably more relevant. Furthermore, Perplexity AI enriched the user experience by offering concise summaries for each reference.

Consider a more complex query: 'Based on the following scenario, find the most relevant academic reference: Recognizing the emotional state of a crowd for law enforcement and security by the combination of visual cues such as facial expressions and body gestures, and audio cues such as vocal intensity, speech rate, pitch, and tonal changes.' While Perplexity AI retrieves 19 references, Bing manages only 3. Notably, the references from Perplexity AI are of higher relevance compared to those from Bing. This highlights the challenge of sourcing such specific references from traditional search engines like Google Scholar.

It is worth noting that the output of LLMs needs to be validated because they are prone to generate plausible-looking incorrect output – an inherent issue called 'hallucination' in

LLMs. The providers of LLMs, such as OpenAI, are currently working on solving this issue (Field 2023).

Regularization of FMs

As we continue to witness advancements in FMs within the AI landscape, the need for policy regulation of their utilization grows more pressing. The European Parliament has made strides in this direction by proposing an initial draft of the AI Act. This regulation aims to establish human oversight of AI systems, promoting safety, transparency, traceability, non-discrimination, and environmental sustainability. Specifically, the proposed Act requires the providers of FMs to ensure the robust protection of basic human rights, health and safety, environmental integrity, and the principles of democracy and the rule of law. This includes the assessment and mitigation of risks, compliance with design, information, and environmental standards, as well as mandatory registration in the EU's AI database (EU Parliament 2023).

However, this proposed AI Act is not without its critics. Many European firms have voiced concerns that the stringent regulatory framework could discourage AI providers, leading them to retreat from the European market altogether (CNN 2023). This could potentially cause ripple effects, compelling various corporations and academic institutions to discontinue their use of large language models (LLMs) in order to align with EU regulations. Such an outcome may inevitably lead to depriving countless individuals and entities of harnessing the extensive benefits of these transformative tools.

A significant concern regarding FMs pertains to their "black box" nature, which severely impedes transparency. This aspect of transparency is a cornerstone of the proposed European AI Act. Enhancing transparency would necessitate unveiling details about data collection, training methodologies, model size, copyright policies, data sources, and an explicit articulation of capabilities and limitations. From a business standpoint, such disclosure may pose considerable challenges for FM providers.

Given these circumstances, one might presume that open-source models could offer greater transparency due to the absence of commercial competition. Nonetheless, recent studies (Bommasani et al. 2023) reveal a different picture; even open-source foundational models fail to meet the transparency standards set by the European AI Act. Certain open-source models like Hugging Face's Bloom rank relatively high in terms of EU compliance, yet they don't reach the optimal level of transparency. In contrast, the widely acclaimed GPT-4 model lags significantly in conforming to the EU Act. Should the EU's AI Act proposal be approved, the future of AI models in Europe hangs in the balance.

Undoubtedly, establishing regulations for LLMs such as ChatGPT is crucial, not just for industry, but also for research and educational institutions. While these entities may not seek the same degree of transparency required by industry regulators, it is vital to implement policies to govern FM/LLM usage within academic and research environments.

It's important that these institutions craft tailored policies for various stakeholders, including educators, students, researchers, and administrators. For instance, guidelines could detail how educators might utilize ChatGPT in lecture preparation or for sourcing

relevant educational materials. For students, there should be clarity on leveraging ChatGPT for assignments and projects in a way that maintains original thought and creativity.

Research powered by these AI tools also presents unique ethical considerations, such as citation protocols, referencing, authenticity, and authorship. These issues warrant dedicated policies to ensure research integrity.

While training and awareness initiatives are essential to maximize the benefits of the expanding AI landscape, there is an equal need for well-thought-out usage guidelines. These will serve to prevent misuse and encourage ethical and productive interactions with these revolutionary AI models.

References

Bommasani R., Klyman, K., Zhang, D. & Liang, P. 2023. Do Foundation Model Providers Comply with the EU AI Act?. Accessed: 04 August 2023

CNN. 30 June 2023. Serious concerns: Top companies raise alarm over Europe's proposed AI law. Accessed: 04 August 2023

CSET. 12 May 2023. What Are Generative AI, Large Language Models, and Foundation Models? Accessed: 04 August 2023.

Field H. 31 May 2023. OpenAI is pursuing a new way to fight AI "hallucinations". Accessed: 04 August 2023.

Ibrahim H. June 2023. Perplexity AI vs. Bing Chat: Which AI Search Engine Is Better? Accessed: 08 August 2023.

Naz H. 2023. Which AI Chatbot is the Winner: ChatGPT Plus or Perplexity? Accessed: 04 August 2023.

EU Parliament. 11 May 2023. AI Act: a step closer to the first rules on Artificial Intelligence. Accessed: 04 August 2023.

Vastola J. 8 May 2023. The Great AI Chatbot Debate: ChatGPT Plus vs. Perplexity. Accessed: 04 August 2023.

Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. Adv Neural Inf Process Syst. ArXiv. /abs/1706.03762.

Zapier. 27 July 2023. The 4 best chatbot builders in 2023. Accessed: 04 August 2023.

Zhang. Y., Gong, K., Zhang, K. et al. 2023. Meta-Transformer: A Unified Framework for Multimodal Learning. arXiv preprint arXiv:2307.10802

Ye, Jiabo, et al. 2023. mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding. arXiv preprint arXiv:2307.02499 (2023).