AI Industry Update

29 March 2024

I mentioned in last week's update that I was going to work on a format and also clean up, somewhat, the content sections. I also said that I'd cover the following topics this week: Agents and Agent Frameworks, Nvidia's AI offerings, Mixture of Experts, and Chain of Thought (CoT)/Tree of Thought (ToT). To those, I'm going to add a narrative of many "popular" Large Models and the companies that are behind them.

One of the reasons that I think we should track the large models is that their capabilities, wide and varied as they are, are inherited attributes of "agents" that act as a layer, of sorts, on top of these models. With this as an introduction I'll start with a survey of the large models.

Before we begin, just a couple of comments, generally, on the current pace and progress the industry is making and what luminaries in this space are saying and or doing.

I thought that 2023 was a year of exceptional innovation and surpassed all previous years of innovation that I have seen in my career. I had anticipated that 2023 was going to seen as the year of innovation and that 2024 would, largely, be a year dominated by implementation and productization of these innovations. I was wrong. 2024 is certainly going to be a year of productization, business and consumer access to these new capabilities, and dramatically increased pace of advancements for efficiency and automation. However, 2024 now looks to be a year full of more innovations than those that 2023 had. Sam Altman, CEO of OpenAI, posted on X/Twitter on March 17:

He certainly isn't one to hold back the hyperbole. The fact that this is all so fast moving and with large advancements happening nearly every week might lead some to say, "well, let's just wait until the dust settles," or words to that effect. It is my strong opinion that there are efficiencies to be gained in real and big dollars, time or pace improvement, and new capabilities. More on that at a later date, perhaps.

Today's update will go as follows:

- Large Model Survey
- Mixture of Experts: a new approach to Large Model architecture
- AI Agents and Agent Frameworks
- Self-reflective large model response models: Chain of Thought and Tree of Thoughts
- Nvidia Advancements

by Bryan Sparks

Large Model Survey

This is just a "fly over" on those models (companies) that are vying for "top spot" in this very large and clearly influential market. I include these so you will be aware of them, their names, and some of their current distinguishing features.

In general, though a particular model may claim "best" on an array of benchmarks, at some point in time it will likely be passed by another in a few months on "their" next release. Nonetheless, these all represent an advancement curve that is stunningly steep. We all benefit.

These newer models are or becoming multi-modal, meaning that they understand language (e.g., ChatGPT) while also understanding or generating images, for instance.

These models have varying capabilities with multiple languages with strengths in one or many.

Also, just to comment on FamilySearch and how we might benefit by these individual advancements. Say, we develop an "agent," or "swarm of agents" that are based on a particular model, generally, we aren't "tied" to that model and could simply change to another and gain additional functionality without changing the agent itself; of course, there are some details circumstances that complicate this somewhat. Lastly, on benefits of several models, again say we have a swarm of agents, some of the agents in the swarm might be based on some model (GPT-4) and others based on another (Mistral 2) yet they all work seamlessly together. So, an agent in an agent mesh can utilize model advancements to complement their defined action while other agents do similarly with other models.

These are provided in no particular order.

1 OpenAI's GPT-4

Brief

As this is the dominant model within FamilySearch and with OpenAI being "first mover" gaining substantial market presence at its launch, I'll say the least about it. It is good, capable on many fronts, and for much of the last year has been the standard whereby others are measured. Their dominance in performance is no longer the case today (Claude 3, see below, is the current leader in most all benchmarks with Mistral's recent release also passing it) and may not be model to be the standard as this year progresses.

TL;DR

One of the historical trends that market leaders have historically done, which has continued with OpenAI, is introduce specific API interfaces to more fully "tie" their customers and developers to their platform. OpenAI's Assistants API, their initial function calling apparatus, and the "OpenAI GPT" user interface are three such efforts. It seems unlikely these will lock people to OpenAI, solely. Other models offer similar features now and there are already popular "abstraction" layers that hide the OpenAI specifics and map to others model's capabilities.

2 Anthropic's Claude 3

Brief

Mentioned in last weeks Update so little is given here. It is very good, includes a Mixture of Experts architecture (see next section of this Update), has a very large memory to retain conversation threads or for large document upload and analysis. It has attributes that will make agentic systems more capable generally and capable of critical thought and reasoning.

It is focused (marketed) to large enterprises and is more expensive to use than all others.

3 Mistral's Mistral 2 and Mixtral (last November)

Brief

A very good and capable model. Latest release (this week) is likely to surpass OpenAI GPT-4 in most all categories. Will be available via Microsoft Azure due to a recent $2.1b investment Microsoft made.

Their commercial (large) model is very inexpensive relative to others.

TL;DR

Mistral is a french startup and made waves in this market last summer culminating with a major advancement in Large Model capabilities last November with their Mixtral release. Mixtral was the first Large Model release that utilized a Mixture of Experts (see below) which showed substantial performance (response value and timing), better use of hardware, and promise in many areas. I believe that ALL future models from all vendors will utilize a Mixture of Experts approach and is within the Claude 3 and Mistral's very recent (earlier this week) of Mistral 2. The next release of OpenAI's model will also include a Mixture of Experts with some rumors that the currrent update of GPT-4 already does.

Mistral has demonstated a shockingly fast pace on productization of advancements without sacrificing stability.

Mistral is a popular choice for fine-tuning as demonstrated on **HuggingFace.co**, where there are currently 331 fine-tuned Mistral models.

Mistral is a commercial concern but has released their model under a license that allows for its use privately. I have this running locally on my personal MacBook Air and have used this on temporary and "rented" GPY systems on **RunPod**. Should an organization desire to use their largest model that they offer from their "cloud" data centers, they are the cheapest alternative to all the others, by far.

Last month (February), **Microsoft invested $2.1b** into Mistral and formed a partnership where their models will be available within Microsoft Azure.

4 Google Gemini

Brief

Not a consideration for FamilySearch at this time. Google is way behind in the "AI wars" and having significant struggles trying to catch up.

TL;DR

Google's Gemini 1.5 release was highly touted by the company and was released just last month (February 15). It was disastrous and, frankly, couldn't have gone worse. It was a major, catastrophic failure for which Google many not recover and if they do recover it will take a long time.

So much of these models are based on an inherit and implicit trust that the system will be truthful and without bias. Though, I acknowledge that I, and others, understand that inherit bias "comes with the territory" on web search (prejudiced ranking) and with chat responses with large language models but the display of the depth of the bias of the Gemini model will be very difficult to overcome. The large loss of "trust" with Google and with any future model from Google, will be sorely tainted with this debacle. Google has apologized multiple times with explicit statements from their CEO. They have also pulled features of Gemini from the market for some indeterminate period of time. Nonetheless, this all may not be enough. Some very bright "market makers" from the "All In" podcast. among others, make this point. It is worth listening, if you have interest, for 15 minutes starting at **this point** in their podcast speaking about Google and Gemini to get a feel on this.

Just one last comment on Google; they invented GPT (Generative Pre-trained Transformer) with the seminal paper entitled **"Attention is All You Need"** in 2017, with the first implementation from the paper, GPT-1, developed by OpenAI in June of 2018. Google had missed the industry-moving importance contained in their own research and paper and continued to sleep for several years and now are frantically trying to catch up.

In spite of all this, Samsung signed an agreement with Google on the smallest Gemini model for inclusion on Samsung phones in January 2024. I wonder if they regret that a bit now.

And just this week, Apple is rumored to be in discussions with Google on the same topic. I'll be surprised if they follow through with this at this time and wonder if this was a Google leak.

5 Meta/Facebook Llama 2

Brief

Not a model ready for productization at this time. Meta's commitment to open source may ultimately win out when this and the other mainstream models achieve AGI (Artificial General Intellegence; sometimes defined as a model that can do 50% of work that humans currently do). Their commitment to open source AGI may make it more interesting but now there are just more capable models to use, play with, be represented as agents.

TL;DR

Llama can be run in many environments which is a plus for developers and those that are "playing" with LLM tech. This open source nature also has made Llama a favorite for fine-tuning of their model (see the many fine-tuned models on **HuggingFace.co**, where there are currently 694 fine-tuned Llama models). This aspect may make this more interesting should fine-tuning, domain specific models become necessary.

6 Grok

Brief

Just released and open-sourced last week. There isn't enough information to make any determinations yet.

Mixture of Experts

Brief

As **Mistral** was the first company/model to do a Mixture of Experts model. which they called Mixtral 8×7b at the time (8 experts within the model each of 7b parameter size), I asked Mistral.ai what the definition of MoE was and this is the response:

Mixture of Experts (MoE) is a machine learning technique used for modeling complex distributions or functions. It's a type of ensemble learning method, which combines the predictions of multiple expert models to improve overall performance.

In a Mixture of Experts model, there are typically two components:

a. *The experts: These are individual models that specialize in different parts of the input space. Each expert is responsible for making predictions in its area of expertise.*
b. *The gating network (or gating function): This is a model that decides which expert to use for a given input. It assigns a weight or probability to each expert, indicating its relevance for the current input.*

The final prediction is a combination of the experts' predictions, weighted by the gating network's output. This allows the model to adapt to different regions of the input space and make more accurate predictions.

This technique could be utilized in models that FamilySearch builds to give better responses, in some cases. If not, then any external model used from one of the major vendors will have employed this technique within their models so a FamilySearch request will benefit from this technique implicitly.

AI Agents and Agent Frameworks

An AI agent is an autonomous software program that can interact with its environment in order to achieve specific goals. A collection of agents can be used to solve complex problems by breaking them down into smaller, more manageable sub-problems that can be tackled in parallel.

1 Unsupervised Learning Breakthroughs

Researchers have developed new unsupervised learning algorithms that can extract meaningful patterns and insights from vast, unstructured datasets without the need for extensive human labeling and annotation.

2 Multimodal AI Integration

Advances in the integration of different AI modalities, such as computer vision, natural language processing, and speech recognition, are enabling the development of more versatile and intelligent AI systems.

3 Explainable AI Techniques

Researchers are making progress in developing AI systems that can provide clear and transparent explanations for their decision-making processes, addressing the growing need for interpretability and accountability in AI applications.

AI Adoption and Implementation

As AI technology continues to evolve and mature, the adoption and implementation of AI in various industries and sectors is becoming increasingly important. This section explores the challenges, best practices, and success stories associated with the real-world deployment of AI systems, providing insights and guidance for organizations looking to harness the power of this transformative technology.

1

Assessing AI Readiness

Organizations must carefully evaluate their data, infrastructure, and talent to ensure they have the necessary foundations in place to effectively implement and leverage AI technology.

2

Responsible AI Deployment

Ethical considerations, data privacy, and transparency must be prioritized when deploying AI systems to ensure they are being used in a responsible and accountable manner.

3

Upskilling and Talent Development

Investing in the training and development of employees to acquire AI-related skills is crucial for successful AI implementation and integration across the organization.