AI Industry Update

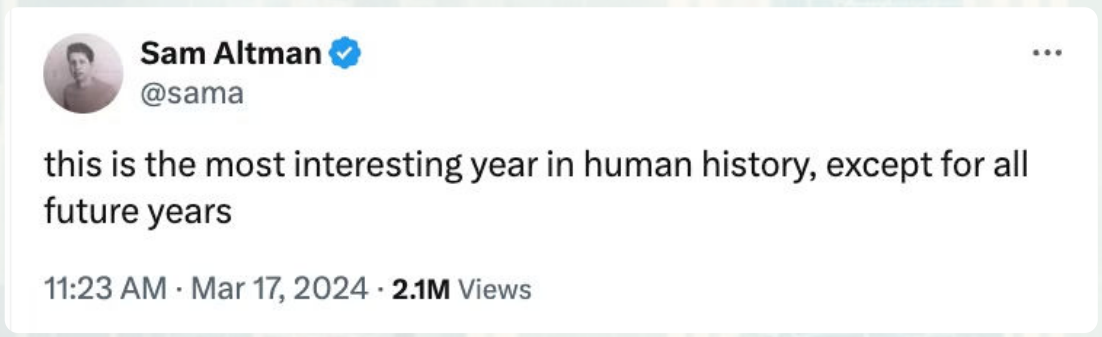
29 March 2024

I mentioned in last week's update that I was going to work on a format and also clean up, somewhat, the content sections. For the sections below I've added a "Brief" and a "TL;DR" section. 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. I'm going to include in this Large Model survey some initial thoughts on Nvidia so I won't have a separate section for that.

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.

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 now, with time or pace improvement, and with new capabilities. Further, learning how an advancement "moves" in time gives a bitter feel for its maturity and readiness for inclusion in our strategies. 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 techniques: Chain of Thought and Tree of Thoughts

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

These models have varying capabilities with multiple languages with strengths in one or many. This multi-lingual feature deserves an update all on it own as we, FamilySearch, is so focused on many, many languages. We'll revisit this at a later date.

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 large models are listed by what I think the short-term influence they will have in the market; this ordering could all change in month or six. The last entry in this list is a collection of "others" with brief comments about them. Some of them I'm a huge fan of for agents that have a define, narrow, and specific purpose that can/needs to be small and run anywhere.

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.

NOTE: Amazon just dropped another $2.75B investment into Anthropic. This brings **Amazon's investment into Anthropic to $4B**.

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 current 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 Nvidia

Brief (long Brief, I know, hang in there 😀)

There is so much to talk about Nvidia that it's tough to summarize. Having near infinite money (**$12.9B Net Income from last quarter**) from their hardware sales they are investing heavily in areas outside of hardware. These other areas include their Foundation Agent, robotics, simulation worlds, agent skill acquisition, agent learning schemes, and many others.

Some of this is probably worth an update at another time because these efforts are changing the way we think of AI Agents.

For today, I'll include a link to a video that is Dr. Jim Fan's (Nvidia AI Research Scientist) **TedTalk**. Don't be fooled that this is only about robotics. It is, I know, because that's shiny and compelling, but it isn't, as well, and instead is much, much bigger. Insights from these capabilities WILL (can) help us find new pathways, products, and efficiencies at FamilySearch, if we choose to investigate and ideate how they could be applicable and on what paths; ignoring the robotics wrapper of this video which is compelling, certainly. Think of the underlying learnings, simulations, models, reinforcement schemes, playgrounds, and the like, as you watch this. It's worth the 10 minutes. Promise.

At another time, I'll review two of their papers from the past few months, **Voyager** and **Eureka**, and how they might help us as we set a strategic path to our own agents, models, and development strategies.

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 Intelligence; 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.

7 Google Qemini

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 to, if you have interest, for 20 minutes starting at **this point** in their podcast speaking about Google and Gemini to get a feel on blunder and the stakes involved.

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.

8 Others

TL;DR

DBRX — DataBricks model. Mostly unknown to me but will be on Azure, AWS, and others. I mention it because DataBricks is a very large company.

xGen — SalesForce's model. Again, mostly unknown but SalesForce is a very large company so could, in the future, make a difference.

Several OpenSource models —

Phi-2 — A small Microsoft LLM with the latest release described from **this link**: "We are now releasing **Phi-2(opens in new tab)**, a 2.7 billion-parameter language model that demonstrates outstanding reasoning and language understanding capabilities, showcasing state-of-the-art performance among base language models with less than 13 billion parameters. On complex benchmarks Phi-2 matches or outperforms models up to 25x larger, thanks to new innovations in model scaling and training data curation."

I'm a huge fan of Phi-2 as a basis for Agents and for fine-tuning. I think Phi-2 represents the future on small, synthetic-data trained LLMs.

BLOOM — Launched in 2022 following a year-long collaborative project with volunteers from 70+ countries and researchers from Hugging Face,

Falcon LLM — Released by the Technology Innovation Institute of the United Arab Emirates in September 2023, their Falcon 180B is being trained on 180 billion parameters and 3.5 trillion tokens. With this impressive computing power, Falcon 180B has already outperformed LLaMA 2 and GPT-3.5 in various NLP tasks, and Hugging Face suggests it can rival models in some ways.

9 This heading is simply an apology for this section being so long.

Mixture of Experts

Brief

This technique, described below, could be utilized in models that FamilySearch builds to give better responses and have higher performance (e.g, less hardware than a single large model). The vast majority of large models we'll see from here forward will have inside a Mixture of Experts.

TL;DR

As **Mistral** was the first company/model to do use a Mixture of Experts technique. 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.

AI Agents, Swarms, and Builders

This section will be a recurring section of this Update newsletter. For today, I'll stay somewhat high level and define the space and some concepts. In future Updates we'll explore specifics, frameworks for construction, inter-agent communication and with humans and what that means, their potential for autonomous (self-guided) functions, and others.

To me, this (agents, agent swarms, agent frameworks, agent builders, etc.) is the most important advancement utilizing Large Language Models (LLMs) of the past year and will eclipse all other advancements (except potentially robots; which ironically is a set of agents captured by some hardware) in the coming year and for a few years to come. We are already where agents and their underlying LLMs are not only understanding language (or image or video), consuming and generation of said, but are now work platforms using LLMs as or converting them to Large Action Models, or what Nvidia is calling a Foundation Agents. Agents and agent swarms can act outside of themselves in whole or in collaboration with other agents and interact with the Web, walk past Captchas, find and use published APIs, or interact with databases, or applications we commonly use; cloud or otherwise. These agents are already capable of:

- downloading apps,
- development libraries,
- generating marketing materials, scripts and accompanying videos and audio for marketing deliverables,
- doing enhanced voice-based customer support for a growing list of organizations,
- acting as software developers and software testers,
- analyzing data from from spreadsheets or databases,
- summarizing complex ideas from published papers, books, or videos,
- capable of doing research on competitors or partners and providing consumable reports of same,
- and so many more.

This isn't future. These examples are today, really, better next month and better and more capable the month thereafter. Their use or initial setup is a bit unpolished, at present, but this will rapidly change in the coming months. Agents and agent networks will transform work across many industries.

Simple Definition

AI agents are given a goal by their creator (real or another agent-builder) and then left to figure out how to achieve it independently using algorithms and other capabilities. These capabilities allow them to understand natural language, access knowledge bases, and exhibit a degree of autonomy. The scope and fidelity of the goal determine how useful the agent can be. More open-ended goals require more advanced agents.

AI agents possess certain key capabilities that are inherited from the underlying Large Model allowing them to function:

- **Natural Language Processing —** Agents can comprehend goals, instructions, and information communicated through human language. For example, an agent could analyze a product support ticket written in English and determine required actions.
- **Knowledge Access —** Agents can utilize databases, knowledge graphs and other data repositories to obtain useful data related to completing assigned tasks. Prior to responding to the support ticket, for instance, an agent may gather details about the related product from an internal support database like Jira.
- **Independent Operation —** Within defined constraints, agents can execute multi-step goals without human oversight. After assessing the support ticket and necessary context, for example, the agent could identify and implement the solution steps automatically take them.

I could go on but will stop for today. In the coming Updates I'll provide real-world examples, videos examining these agents and networks, and provide potential candidate frameworks (building systems) that we at FamilySearch could use to explore them.

Lastly, to restate what I said at the beginning, this is much larger and, I firmly a strategy we should explore, to fully leverage the wave of explosive innovation and AI models and their growing capabilities of the past year; agents and agent networks is a bigger idea than chatbots, RAG (Retrieval Augmented Generation) systems, and all the others we could mention as each of these are, you guessed it, potential agents acting individually or in a larger "swarm" or mesh solving some higher level task given it.