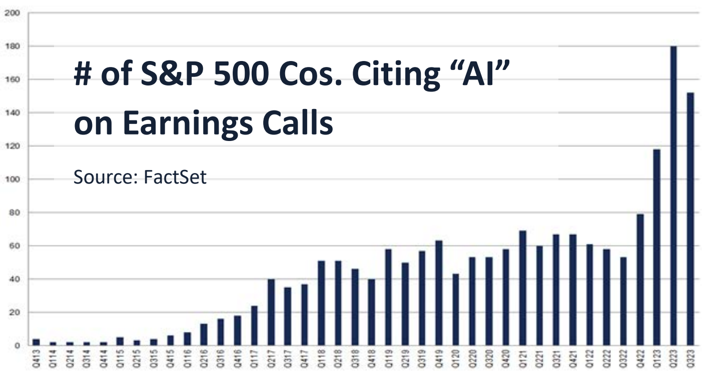
AI Outlook: Future Applications & Limits of LLMs

FEBRUARY 2024

Ever since the explosion of usage and media about OpenAI's ChatGPT in March 2023, companies across all industries have raced to roll out "AI strategies" as quickly as their next earnings calls.



Evidenced by the immediate surge in Nvidia's GPU sales, few initiatives were as high a priority at many companies. The promises and potential of AI are vast, from unique customer-facing products to a significant reduction in customer service costs. The threat of falling behind may be an even bigger motivator, as companies from Chegg to Adobe saw investors panic about the real or imagined impact of AI. As we approach the first anniversary of ChatGPT's breakout, we find it worth asking how far Large Language Models (LLMs) have come in the past year and where they may (or may not) be headed.

Basic recap of how LLMs work and why ChatGPT is so powerful for writing

Foundation Models, or Large Language Models when referring to text-oriented models like ChatGPT, work by tracking relationships in sequential data (we will refer to both as LLMs as the underlying technology is similar). This includes data streams that human brains process every day – pixels in an image, notes in a song, words in a sentence – or data people are not good at directly interpreting such as computer code or protein structures. LLMs use 'transformers', a concept invented by Google, which allows for 'unsupervised' learning. The models apply positional tags in data coming into the model, which creates an increasingly complex understanding of the relationships between existing data and new data inputted into the model. While prior neural networks required labeled datasets to work effectively, transformer models can ingest huge amounts of unlabeled data and develop an understanding of context and patterns mathematically. The results are significantly more powerful, with highly accurate interpretation of complex language inputs. The catch is the computing power (and thus cost) required to train the latest LLMs is orders of magnitude greater than previous neural nets given the vastly larger datasets being processed. This has given way to a relatively small number of popular "pre-trained" models, including multimodal models capable of different types of input such as OpenAI's GPT, or Google's Gemini, or text-oriented models such as Meta's Llama2, or Anthropic's Claude2. These can be further refined by companies with proprietary datasets to try and develop more accurate models for specific use cases.

Limitations of LLMs

These models have a probabilistic, not deterministic, understanding of context, which has important ramifications for their power and limitations. This is why early articles covering ChatGPT pointed out the model could instantly spit out creative lyrics but struggled with basic math. Minor tweaks in input can lead to a wide range of output, as the context a user provides the model impacts the probabilities it assigns to the most likely data to output. This is a good thing for many use cases like a customer service chatbot or music composition tool. A frustrated customer can make progress with a back-and-forth conversation instead of getting stuck with the same canned output of a deterministic program. An artist can tweak and iterate on what sounds are input and get different outputs each time. It is also a bug for many use cases. When attempting to use ChatGPT instead of Google Search, results can have a Jekyll and Hyde quality – sometimes comprehensive and informative, while sometimes confidently stating made up information. The need for manual verification is a significant difference between the automation LLMs can enable, and something like a calculator (or more complex implementation like CAD software). One can easily infer the potential to come up with novel molecules in drug discovery, but also see that human review and investigation is still necessary.

The Promise of "General Intelligence"

The path to a truly "general" intelligence model relies on significant innovations in technology that have not yet been cracked. Though there is no precise definition for "general" AI, for investment purposes I believe it is fair to set the bar at meeting or surpassing human performance in complex, cross disciplinary tasks. First, LLMs consistently score poorly on abstract reasoning and logic tests, a critical measure of human intelligence and key to solving complex problems. Second, they are not effective at generalizing patterns or learning approaches from a dataset they have trained on to solve a novel problem, limiting their usefulness for 'interdisciplinary' work. Third, hallucination, or a lack of precision in general, limits applications for some areas.

Vinod Khosla and Geoffrey Hinton both predicted years ago that radiologists would be obsolete by 2023, yet current headlines suggest there is a shortage. This viewpoint reflects a caricaturization of many white collar professions; in this case the radiologist staring at an obviously broken leg or cancerous lung. Every investor that has by now heard how AI will be able to find 'cheap' stocks should sympathize. In reality, most jobs are highly specialized and involve far more process and complexity than basic image recognition or data entry. This is why we have seen earlier predictions of mass unemployment shift into a litany of 'copilots' to make still-employed office workers more efficient. AI is simply not precise or consistent enough to replace a radiologist's diagnosis, which relies on more than a quick glance at an image. However, a specialized model could certainly help with measurement and identification to a degree that makes human diagnoses and recommendations even more accurate.

Small improvements have been made in all of these areas of weakness, and a major breakthrough may be forthcoming – though that is not something I have the expertise to predict. Given the recent impact of AI on the market, it is vital to understand that most of the current spending on AI is not even attempting to achieve this outcome. Hype cycles are common in technology, and we have seen expectations from both the media and the market skyrocket in the past year.

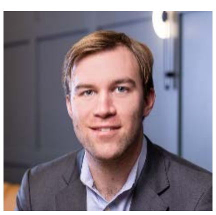
Real Implementations of AI Models

While ChatGPT has been a hit for students needing to crank out a last-minute essay, it has not yet transformed the lives of knowledge workers. As mentioned previously, most enterprise spending on AI is meant to make models more relevant to their specific industries and organizations. This is typically done by customizing a 'standard' pre-trained model from OpenAI, Anthropic, or Meta's open source Llama2. Due to the hundreds of millions of dollars required to train such massive models from scratch, it is far cheaper and less risky to instead add a custom, proprietary dataset and tweak the model to focus more effectively on a specific type of query. For example, call center chatbots can be trained on transcripts of past interactions to focus on specific customer service issues with greater precision, and cut off open ended responses involving irrelevant topics.

Obvious winners here are the few large model providers, compute infrastructure vendors (both cloud providers and the producers of the specialized GPUs required), and large companies with valuable proprietary datasets and the ability to invest in R&D. This approach is a direct path to productivity improvement and value creation for companies that do it well. It will not lead to a great leap forward such as a model that can truly reason or learn wholly new tasks with no retraining. The outcome will be a meaningful but not revolutionary wave of copilot-like products that deliver enough productivity uplift to command value for software vendors. Microsoft has been the most public with this effort, but every large software, services, and data vendor is looking at a similar approach.

Conclusion

The copilot wave is driving a boom in infrastructure spending to support so much sudden interest. I do not believe this will lead to enormous disappointment and a collapse in spending, as the applications are diverse enough to allow for plenty of failures. However, without more significant leaps forward in capabilities, I expect AI spending will plateau as companies decide it is more evolutionary than revolutionary. In our view, the longer-term winners will be existing category leaders with ownership of industry datasets or control over customer workflows who can augment their products to offer more value or their processes to reduce costs. Many hardware vendors will see margins squeezed as growth slows and large cloud providers focus on cost optimization over footprint growth. A real breakthrough would alter that path and perhaps drive the AI boom into a new level of mania.



Brian Tolles
Lead PM, Global Growth
Jackson Square Partners

AI Around the Globe

The AI arms race has further propelled US equities past global peers given the preponderance of large US technology companies in the value chain. While the "usual suspects" in semiconductor production are clear non-US beneficiaries (e.g., ASML, TSMC, and smaller suppliers such as BE Semiconductor), governments and investors around the world are eyeing the opportunity to encroach on the dominance of US tech. The EU, UAE and China have all committed varying degrees of state capital to local AI companies. A few European startups, such as Mistral in France and Stability AI in the UK, have gained real momentum. The EU and China have also been most aggressive in rolling out AI regulations addressing issues like copyright, privacy, and model transparency. However, AI is a far more difficult invention to regulate than more concrete advances such as self-driving cars. The term "AI" does not have any agreed upon definition, with no clear-cut line that makes one piece of software 'intelligent' and another not. Furthermore, trying to stave off potential issues with regulation instead of regulating externalities after the fact will impact the development of a region's corporate ecosystem. Large US tech companies have arguably benefitted most from the cost and complexity of EU regulations like GDPR. One of the biggest advantages startups have in the AI arms race is the lack of a reputation to lose – moving fast and shipping is more important than accidental copyright issues or 'AI safety' concerns. The cost and delay from dealing with investigations or potential fines could significantly reduce this advantage. So far, Big Tech looks well positioned to continue its dominance in these markets and emerging startups seem like marginally important companies.