What Was I Made For LLM

Ubiqus
61 Broadway – Suite 1400 – New York, NY 10006
Phone: 212-346-6666 ◆ Fax: 888-412-3655

What was I made for LLM

[START RECORDING]

FEMALE VOICE 1: This podcast have been prepared exclusively for institutional wholesale professional clients and qualified investors only, as defined by local laws and regulations. Please read other important information which can be found on the link at the end of the podcast episode.

[Music]

MR. MICHAEL CEMBALEST: Hello, everybody. Welcome to the September Eye on the Market video audio podcast. This one's entitled "What Was I Made For: Large Language Models in the Real World." I wanted to focus on this topic again because of how large AI is as a catalyst, what's going on in the equity markets. But first, I just wanted to review economics and market for a minute. Not that much has changed since our August piece called "The Rasputin Effect."

Leading indicators are definitely pointing to weaker growth by the first quarter, but the expected decline is pretty modest as potential recessions go. Tighter credit conditions are certainly going to have an impact, but only 17 or 18 leading indicators that we watch, none of them looks really terrible, they all just look kind of modestly bad, and a little bit weaker.

The reason why things don't look worse after 500 basis points of fed tightening is that the fed policy is being offset by a few things. First of all, very large fiscal deficits, almost as large as they were in 2009. We're having the beginning of a US industrial policy which is essentially incentive-driven spending by the private sector on infrastructure, energy, and semiconductors. That's starting to kick in, but household and corporate balance sheets were pretty strong coming into this year.

Delinquency rates outside subprime auto are still very low. The private sector took actions to lock in low borrowing rates before 2022. Apparently the only entities that didn't get the memo that rates were unsustainably low were a handful of some of the regional banks that you're all familiar with who extended their asset duration at the wrong time.

Housing markets and labor markets are pretty tight, so the normal transmission of higher interest rates and higher fed policy to crater housing and labor markets isn't transmitting

quite the same way. So, all of these things are, at least at the current time, kind of keeping a severe recession at bay.

I do want to talk a little bit here about oil prices. The OPEC spare capacity that is pretty high, it's not as high as it gets during recessions as you can see in this chart, but it's pretty high. For a non-recessionary period, OPEC has engineered quite a bit of spare capacity. Now, that can change quickly, but right now spare capacity is pretty tight. You have to combine that with two more things.

First, the publicly traded energy companies are spending a very small share of cash flow. We have a chart in Eye on the Market that shows the percentage of energy company cash flow that they're spending on new projects, specifically oil- and gas-related projects, and that's a very low share, and we juxtapose that against global fossil fuel use. You can see the industry is starting to cut back on future projects for all the reasons you might imagine, even though we really haven't see much decline yet in global oil and gas consumption.

Then on top of that you've got the Strategic Petroleum Reserve at the US at the lowest level it's been in many decades. So, tighter OPEC conditions, less oil and gas investment, and the depleted Strategic Petroleum Reserve, that combines to kind of goose up oil and gas prices, and then we'll have to see what Russia has in store for the world. They've already announced some restrictions on diesel exports.

Higher energy prices tend to feed into inflation within a few months, and so one of the things that you're seeing is the markets were pricing in some fed cuts next year; that's now gone. Now, I did want to focus most of this discussion on generative AI catalyst, because we have a chart in the Eye on the Market this time that shows an ETF for generative AI stocks is up around 60% this year while the market, excluding those stocks, is up around 5%, so this has definitely been the year of generative AI.

I wanted to take a look at how it's being used well and where it's failing, and then perform my own specific test on GPT4 specifically, because I thought it was an interesting exercise. The reason I want to do that is juxtapose these two things. Number one, people are out there comparing large language models to electrification of farms, the interstate

highway system, and the internet itself, those are kind of some pretty remarkable milestones.

While at the same time we just lived through a period, whether it was cannabis investing, non-fungible tokens, metaverse, block chain, crypto, hydrogen, where a lot of things were kind of touted to be something that they turned out not to be. So, now we're getting a surge and interest in the large language models, and I think the reality is somewhere in between the nonsense of the metaverse and crypto and the seismic changes introduced by the interstate highway system, and then electrification of farming. So, let's take a closer look.

I started out just doing something lighthearted but still meaningful which is there are these multimodal AI image generation models, and I used three different ones you can see here: Bing, Starry AI, and Dolly, which is GPT's version. I asked it to create an image of two people sitting at the table looking nervously at a robot with them, and that the robot should have a label on it that says "Strategy Team Trainee," like working for me. None of them did it right, and some of them, the mistakes are interesting.

So, starting on the left, first of all, there's three people, not two, and one of the people looks like they're in a horror films, which is pretty scary. Lots of people have extra hands and legs and fingers and things like that. The second one from Starry AI got a little bit closer. You have somebody looking nervously at a robot but there's only one person instead of two, and both the first two ignored the whole thing about the Strategy Team label entirely.

Then you have this Bergmanesque and also fairly terrifying offering from Dolly on the right, splattering some letters on the table, not on the robot, and not really spelling anything. So, I thought this--but still, the interpretative proficiency is good in certain ways, so I thought this mixture of good, bad, and bizarre was a good way of starting this discussion.

Some of you will pick up on the theme of this and the pop culture references I'm using, but when you think about a large language model and something it's made for, here are some examples that are currently working. It's helping management consultants in terms of speed and quality and task completion.

Whether you're impressed with that or not depends on what you think of management consultants. People using Copilot, which is a programming tool, are having a lot of success with it. It's doing a great job on statistics. It's helping people that do professional writing. It's helping customer support agents be more productive. It's improving their employee retention, and a lot of these things tend to help the lower-skilled workers the most. It's even having some successes in medical research.

The one that I thought was interesting, where somebody fed in some of the 70 most notoriously difficult-to-diagnose medical cases just based on the descriptions of the symptoms people were having, and it got two-thirds of the diagnoses correct. Now, you're not going to like all these large language model use cases. People are using them to generate digital mountains of thick content, fake news sites, fake product reviews on Amazon, fake e-books, phishing emails--I spelled phishing wrong because I like fishing so much--I should have spelled it with a P-H.

A lot of this stuff seems designed to profit from Google, essentially fool Google's automated advertising process into paying it for people looking at junk content that they don't really know is AI-generated. In any case, these are the things that it's doing well and where the use cases are expanding.

I saw this chart from Open AI but I wasn't as impressed as I think Open AI wanted me to be. It's a chart that shows how GPT4 is doing versus GPT3.5, taking all sorts of standardized tests. As you can see here, there's math tests and chemistry exams, bar exams, biology exams, history exams, SATs, GREs, things like that.

There's something, I think a lot of you are probably pretty aware of this right now, but there's something called data contamination which is if you train these models on information sets that include the questions and the answers to all these exams, all we're really analyzing is whether or not GPT, or any of the other ones, whether it's Bard or Bing or Anthropic or any of the rest of them, they are good at memorization.

But we know that large language models are good at memorization, so I'm not really sure exactly what's being proven here other than the impact of having 10 times more

parameters in GPT4 than GPT3.5 makes it better at memorization.

I think the more important question is you don't hire a lawyer so that he can sit down and answer bar exam questions all day, you hire a lawyer when you need somebody to integrate new information and evaluate things maybe they haven't seen before. When you look at those kinds of tasks, large language models aren't doing quite as well. We have a page in here called "It's not what I'm made for."

When GPT4 has been asked to take law exams it does pretty poorly, and I like the description from the University of Minnesota professors who did this where they said "GPT4 produced smoothly written answers that failed to spot many important issues, much like a bright student who didn't attend class and hadn't thought deeply about the material."

So, now you can get a better feel for what we're dealing with here. It's like repetition rather than real reasoning and thought. GPT4 did terribly on the actuarial exam, a college sophomore economics exam, graduate-level tax and trust and estates exams. It botched Pythagoras' theorem when being asked to be a math teacher. It got stuck in a death loop of nonsense when somebody provided it with mathematically impossible dimensions of triangle that it should have been able to figure out.

The journal had this article where they're writing about how online editors and newspaper editors are being given so many crappy AI-written submissions that they have good spelling and grammar but lack of coherent story. They're just outright rejecting anything that they can get the sense that there was any AI used to generate it at all.

The most comprehensive assessment of large language models that I've seen is something called Big Bench, which is a project that over 400 researchers around the country are working on. There's 204 tasks involved, and the latest that it was updated was July of 2023, of this year, and they still found substantial underperformance of large language models compared to the average human, much less the highly performing human.

Anyway, Manuela Veloso is from Carnegie Mellon and she runs JP Morgan's AI research group, and they're doing a lot of really interesting applications of large language models. She walked me through some of them and I was very impressed. They do seem like they're productivity savers, information

checking, information gathering, charting tools, making sure that documents are filled out properly, all of which are mostly designed to reduce errors and omissions, and that's potentially a very powerful and profitable application of a large language model.

For me, it's a little different. So, here's what I did. I took 71 questions from the Eye on the Market over the last two years that my analyst and I worked on, and I asked ChatGPT4 to take a shot at it, and I graded GPT4 based on its speed, accuracy, and depth, versus the work that we had done ourselves to get the answers. In other words, we're not grading it whether it can do anything, we're grading it compared to the process that we use that didn't yield and hallucinations or errors or things like that.

We enabled the GPT4 features to upload data files when it couldn't find it on its own and needed date files. We enabled the plug-ins that allow it to browse PDFs and Excel files when necessary. So, as a result, a lot of you have read that GPT4 is training data for its parameters ended in 2021. That's not a constraint because we added all the plug-ins to give it all of the data and all of the web access that it needed to answer any of our questions.

So, here are the results. It was a mixed bag, and a very bimodal distribution of grades. It got a lot of As. Out of 71 questions it got 26 As and 25 A-minuses. That sounds great. The problem is, it also got 13 Ds and 6 Fs, so it was very much of a bimodal distribution. The GPA worked out to around 2.5, which is between a C- and B+. You might say, well, what did it get wrong?

Here are some examples of what it did. It would hallucinate numbers and then absolutely refuse to provide a source for where it found them. It was very frustrating. It would outline the correct steps to solve a problem and then execute the steps incorrectly when doing it. It misread data files that we provided to it. It didn't notice when there was data in a spreadsheet and there were subtotals that you should exclude subtotals from when you're summing a column. It messed up some energy conversions, and it also asserted certain facts that are easily contradicted by other readily available information.

So, that was my experience with it, and I guess the bottom line is, just to wrap up, I think GPT4 is going to have a big impact in Manuela's world, for example, since the tasks that

she's designed for it conform more to what these things are made for, which is error checking and memorization, most often using trained corporate data and not just trained internet data.

The part that I struggle with the most is how am I supposed to incorporate a tool where even if it can get some answers to complex questions right, I have to check every single answer, because since it sometimes gets things wrong I have to check every answer, and by the time I've done that, where's the productivity gain of using the tool in the first place.

So, anyway, I'm just going to use it for the simpler questions where it performs well. I think that's what it's made for, and at just $20 a month for GPT4, I got what I paid for.

So, that's this month's Eye on the Market. We've got a piece coming up that's a deep dive on New York City and its recovery compared to other major metropolitan areas that I think a lot of our clients will be interested in, and of course, we're going to continue to monitor what's going on with the fed and consumer spending, energy prices, and economic slowdown later this year. Thanks for listening, and I'll see everybody next time.

FEMALE VOICE 1: Michael Cembalest's Eye on the Market offers a unique perspective on the economy, current events, markets, and investment portfolios, and is a production of JP Morgan Asset and Wealth Management. Michael Cembalest is the Chairman of Market and Investment Strategy for JP Morgan Asset Management and is one of our most renowned and provocative speakers.

For more information, please subscribe to the Eye on the Market by contacting your JP Morgan representative. If you'd like to hear more please explore episodes on iTunes or on our website. This podcast is intended for informational purposes only and is a communication on behalf of JP Morgan Institutional Investments, Incorporated.

Views many not be suitable for all investors and are not intended as personal investment advice or a solicitation or recommendation. Outlooks and past performance are never guarantees of future results. This is not investment research. Please read other important information which can be found at www.jpmorgan.com/disclaimer-EOTF.

[END RECORDING]