Pathways to the Future:
Polen Perspectives on GenAI, Part 2

Dan Davidowitz and Steve Atkins of Polen's Large Company Growth team join Zhang Zhang, Head of AI and Risk Analytics, for a wide-ranging discussion on the technology and investment implications of generative AI.

Steve Atkins, CFA
Portfolio Manager, Strategist & Analyst
24 years of experience

Dan Davidowitz, CFA
Portfolio Manager & Analyst
25 years of experience

Zhang Zhang, CFA
Head of AI & Risk Analytics
25 years of experience

Q: Agentic AI seems to be the next big focus area on both the enterprise and the consumer side. On the enterprise side, you could have AI agents running in the background of a company, helping manage workflow, accounting departments, and other functions. On the consumer side, agents could be deployed to plan a vacation for your family, manage your schedule, or order a coffee. In both areas, users must trust that the model will work correctly and produce the accurate, desired result. Do you think this trust factor could be a major hurdle to acceptance, particularly for enterprises?

A: My view is that people tend to be skeptical of any type of systematic takeover of something that previously wasn't systematic. Some would argue it's unfair that we subject these systems to essentially a 0% failure rate standard when humans might have, say, a 3% to 5% failure rate. I think the difficulty here is that if there is an agentic failure, it might fail in different ways than a human. These models don't work the same as our intuition.

They work by predicting what a right answer might look like. They don't actually know what the right answer is, but may offer responses that sound perfectly plausible, just like a person who's skilled at making things up in a believable way.

That's dangerous because you don't know until the last moment that something might be far off. By that point, you've got less time to take over and react. Those are real issues, although it's not to say that we can't work through them. These models are becoming more transparent. There are ways to force them to cite what they're doing, to shed light on their reasoning and thought process, so humans can audit and review.

With the prevalence of technology today, I think we've grown accustomed to some level of failure rate. However, I believe different hurdles exist in the consumer and enterprise spaces. For enterprises, it depends heavily on what the task is. If it relates to clients' privacy or security, or protection against fraud, the reliability must be almost perfect. But for other things, like enhancing workers' efficiency rate, companies could tolerate a much lower accuracy rate. It really depends, and I've been surprised at how willing some people have been to play with the technology, even as it has had significant limitations.

Q. What do you see as the primary drivers of these accuracy issues? We've observed how some aspects of these models naturally improve over time, but there's also a data limitation issue. Once the models have ingested all the available data, are you concerned about their potential to create unreliable synthetic data?

A: Yes, it's an age-old problem. I think in some ways, the early versions of GenAI models amplified the issue because there are so many categories of risks. There are the data risks of what it's

trained on, which in many cases are the American internet, publicly available sources, and a lot of potentially copyrighted sources. There are English-language, age-related, and other biases in the model, given how it was trained.

Because the model is operating with a lot less information than you might think, it may indeed give you flawed answers that you have to be very critically aware of when you're using it. I think these problems and the solutions to them are very similar to other problems that we faced for decades in the past, regarding data quality, data availability, and integration. These issues are just as important now, if not more so.

Q: OpenAI has captured significant attention and spent a lot of money creating some of the most advanced models. But then, in January 2025, the news on DeepSeek hit the market.[1] This put China back in the conversation again about how advanced their AI engineers are and the models that they were working on. Also, it brought more attention to the cost of running the models. What reaction did you have to this news?

A: While understanding that there might be some hyperbole in their cost estimates, I thought it was interesting to see the first major indication that there may be more efficient ways to run these models. As we discussed, the information is out there. Everyone has access to it and people can improve on earlier models. The DeepSeek news indicated that we don't necessarily need to keep throwing more chips and hardware at the problem to get good data for at least some workloads, especially those that are inference-based.

Most companies are not producing their own hardware; they are renting it as a service from others. And so, anything that brings down the cost of compute for generative AI and allows for equivalent or better use cases over time is good. Now, we're a little bit biased on this issue from the standpoint of our Focus Growth portfolio, in that we don't own the hardware companies, but we own a lot of the buyers and users of it. So, for us, it would be wonderful if the models ended up becoming more efficient, the purchasing needs declined, and applications could run with less computational intensity.

My observation is that some market watchers tended to take one or two years of sales data and growth rates for the likes of Nvidia and the other hardware manufacturers and almost use a ruler to draw a straight line and extrapolate their growth into the foreseeable future. In my view, this first development from DeepSeek could be one of several that signal the future is less certain than we thought.

In some ways, it's surprising that it took so long for this news to make an impact. In 2024, DeepSeek's Version 2, their pre-reasoning model, had already made some splashes in the development community with practitioners noting that it rivaled some of the best existing models at a fraction of the cost. And, with all the papers DeepSeek publishes, we glean that they've discovered ways to build models with limited memory bandwidth and processing power, which is likely a direct result of the export bans on chips. So, this has spurred innovation, and it's not happening in a vacuum. We've seen Open AI make strides in reducing their costs as well.

From our internal benchmarks that we're running with some of the tasks that we're building, we found it was passing 100% of our test cases. We have harder tests now because GPT 4.0 was passing perhaps 30% of those, and everything's working now for half the price. So, we see this as a longer-term trend line. When something new is invented, in hindsight, the first version always seems awkward and inefficient. But if it remains useful, people will figure out how to take the useful bits out of it.

When we were thinking about potential investment in semiconductor companies and Nvidia, in particular, one big challenge was that we'd have to have a prediction about how big the market for generative AI products and services would be, which is extremely hard to do with new technology. Secondly, we'd also have to be right about the future compute intensity and the associated demand for GPUs and servers. As we are witnessing now, you can't just "copy and paste" what came before.

The big tech companies went full bore in meeting their customers with generative AI because nobody wanted to be left behind. Microsoft had already "hitched their wagon" to OpenAI. Amazon followed suit, and then Google and Meta, giving rise to an "arms race" and a bolus of demand.[2] They were willing to make a massive investment. GPUs are fungible, so these companies would rather overbuy than underbuy because if they're wrong about the demand, they can use them for other things.

With DeepSeek and, I believe, with more innovations to come, there will be questions about whether the current method is the only method. Do large language models need to be that large? We're already seeing them come down for a lot of enterprise applications. In these cases, they are designed to be trained on a company's own datasets, which are smaller and more focused.

Q: Slightly after the DeepSeek news, we saw several hyperscale cloud companies report and give their 2025 guidance for capital expenditure (CapEx), which was very high—exceeding even the loftier market expectations. Clearly, they may be seeing the same efficiencies, but it's not affecting their spend plans. Why is this?

A: I think there are a couple of reasons. The 2025 capital expenditure plans for all these companies were already baked, and their big projects are already in motion. Large data centers can take

two to five years to build. Meta is building a data center almost the size of Manhattan.[3] They're not going to stop it because DeepSeek had a breakthrough.

I think what will be more interesting, as these companies start to think about 2026 and beyond, is whether there are potential efficiencies they can achieve. None of these companies are excited to spend tens of billions of dollars with Nvidia. They would love to have more vendors, operate with less compute intensity, and generate more free cash flow. Some are making their own chips and custom silicon. They're looking for more efficient ways to run their workloads. These initiatives were not going to stop the CapEx freight train that was already in motion for 2025, but they do raise some interesting scenarios for 2026 and beyond.

Q: Recently, the term Jevons Paradox has come back into the lexicon. This says that as the cost of a resource comes down, the utilization of that resource expands substantially. So theoretically, as the cost of AI comes down, the use of AI should expand as well. This would seem like a "good problem" for the cloud companies, right?

A: This is not a paradox, by the way! I think a paradox is something that doesn't seem to make sense. This makes a lot of sense. If something is cheaper, people want more of it. You would hope that if the cost of computing comes down, lots of industrious people will figure out ways to make even better AI products and services because of that. It would be wonderful if that happened!

In my opinion, I think some people are using this so-called paradox to justify their straight line on semiconductors. Effectively, they're saying, "Don't worry. As the cost comes down, we'll make up for it in volume." We'll see. In the long run, that might be right. But it certainly opens an air pocket. We've already seen so much capacity pushed in, and there is a big possibility that we won't need that much for the next few years. Again, it just underscores how uncertain the future is. I think one must be reasonable and reflect that if you thought the solution was a lot of GPUs and a new innovation comes along to challenge that assumption … well, that be the real story. You should be intellectually honest enough to acknowledge that possibility and not just insist that the difference from lower costs will be made up on volume.

polen capital

Footnotes

[1] Reuters, "DeepSeek leaves US AI firms racing to understand its success," January 2025
[2] Microsoft, Amazon, and Alphabet (formerly known as Google) are holdings in Polen's Focus Growth and Global Growth Portfolios as of December 31, 2024. Meta and Nvidia were not holdings in Polen's portfolios as of the same date.
[3] Data Centre Magazine, "Meta's 2GW Data Centre: How the Company Plans to Grow AI," January 2025

Important Disclosures

Going Beyond with Polen Capital

Polen Capital is a team of experienced investment industry professionals who share an unwavering commitment to our clients, investors, community, and each other. We have been dedicated to serving investors by providing concentrated portfolios of what we believe are the highest-quality companies for more than three decades. At Polen Capital, we have built a culture of results, and in this, an inherent belief in going beyond what's expected for the people and communities we serve.

We adhere to a time-tested process of researching and analyzing companies around the globe—seeking only the best to build highly concentrated portfolios. Then, we invest for the long haul and with a business owner's mindset, giving these companies time to grow.

polen capital