AI360 podcast

Season 1, Episode 6: 2024 AI predictions: Navigating a dynamic market

Host:
Rohan Gupta, Principal, Deloitte Consulting

Guest:
Baris Sarer, Principal, TMT, Deloitte Consulting

Rohan Gupta: Hello, everyone. Welcome to a very special episode of AI360. I'm joined by a friend of the show, Baris. Thanks again for joining us.

Baris Sarer: Great to be back here.

Rohan Gupta: Well, we've, we know, we've talked about 2023 in the last few episodes and what some of the major advances are. Now we're going to look at 2024 and make some predictions about how the year in AI is likely to unfold. So without further ado, Baris, maybe you can kick us off with some predictions you have in AI technology.

Baris Sarer: Sure. From a technology standpoint, while I don't expect a step-function change in core capabilities, there will be a shift in focus from chatbots to end-to-end automations. We'll start tapping into the reasoning capabilities of LLMs [large language models] through autonomous agents, and in parallel, we'll start seeing improvements in model architecture that will make them more efficient, requiring fewer parameters and less data to train, and they'll be able to hold more context.

We'll see some changes in modalities as well. They will continue to mature rapidly. In 2024, text image and voice will become commercial-grade, and then video and music will continue to mature but won't be ready for primetime yet. And lastly, I want to talk about perennial-open versus closed-source debate. [We'll] see strong innovation on both sides of the spectrum. We'll see closed commercial, multimodal, very large general purpose models getting increasingly sophisticated in capabilities. And then we'll also see open-source, medium-sized models of, say, 10 to 100 billion parameters and small-size models that would be, let's say, sub-10 billion targeting the broader developer communities, enterprises, and also OEMs who want to embed models in their products, and then will also see more special-purpose train models targeting different industry verticals.

Rohan Gupta: Wow. So tons of development all across the models and the software space. What does this mean for hardware and for silicon providers?

Baris Sarer: Great question. As models get more performant at increasingly smaller sizes, we'll see a shift in AI processing from cloud to edge devices, and that includes PCs and handheld devices. And this trend this year will be further catalyzed by major software vendors' desire to offload more AI processing at the edge. And this will lead to the emergence of a new market for AI-enabled end-user devices. And to give you a sense of the scale of this opportunity, total share of AI-capable PCs to be shipped next year, or this year rather, is about 19% of all expected AI shipments, and this will be an important space to watch in 2024.

Now, more broadly, in terms of infrastructure and AI delivery, competition will intensify in the space. While I don't believe the pecking order will dramatically change in 2024 yet, the shift to edge will create new winners among both chipmakers and OEMs. There will be new entrants to the market to deliver AI processing from the cloud and also the desire to diversify suppliers. Both for risk and commercial reasons—and by the way, this applies to both large companies as well as sovereign nations—will drive more investment in innovation in chipmaking and cloud infrastructure that will be critical to AI delivery.

Rohan Gupta: Yeah, it sounds like 2024 is going to be huge even for hardware and silicon. So if that's what's going to happen at the technology layer, what does this mean for the vendors actually providing these solutions? Any predictions for the vendor landscape in 2024?

Baris Sarer: Well, 2023 actually was a down year. If you look at the overall market, both for VC [venture capital] investments and M&A [mergers and acquisitions] activity, and despite that fact, we've seen 15 new AI unicorns emerge last year. And according to Crunchbase, more than $1 in $4 invested in American startups in 2023 has gone to an AI startup, which is twice as much the average of the preceding five years. And this level of investment will have strong implications for 2024.

We expect AI investments and growth in valuation to continue into 2024 and perhaps even get turbocharged, especially if the overall market conditions improve. And then as we see demonstrable business value delivered by AI. And under the circumstances, we'll likely see new market entrants, some more unicorns, definitely larger deal sizes, especially with late-stage mega rounds, and in general more M&A activity overall.

And then I want to make a couple additional predictions on the vendor landscape. We will see some new software vendors achieve significant growth and valuation and ARRs [annual recurring revenue] crossing the billion-dollar mark. And also, last but not least, hyperscalers will hold on to their LLM-agnostic position, whether they offer their own models or not, and continue to keep the eye on the real prize, which is, for them, cloud consumption.

Rohan Gupta: Makes sense. So it sounds like actually the economics are going to start turning a little bit more favorably for the vendors in '24. So let's talk about the actual buyers, you know, enterprises in this case. What is this going to mean for adoption for enterprises? Yes, the technology is going to mature. Yes, the vendors are going to make more money. Are enterprises going to be adopting more AI as well?

Baris Sarer: Well, in 2023 we've seen many Gen AI products come in commercially available, but we have not seen much in the way of Gen AI-enabled solutions getting deployed in the enterprise. That's the first thing that will change in 2024. We'll see a steady stream of copilot use cases going into production and business value realized. The economics of this process, as we just talked about, will improve thanks to the intense competition in the AI infrastructure space. As far as autonomous agents will be in the early innings in 2024, I don't expect large-scale automations yet. But we'll definitely see early examples in the form of proof of concepts and pilots.

And lastly, in terms of LLM adoption, this will not be a winner-takes-all market. We will see multiple models in use, both in enterprise and consumer applications, from open to closed source, and from very large to small models. This is not to say that all models will be created equal either. There will continue to be some market-leading models, both in the closed- and open-source space.

Rohan Gupta: Yeah, makes sense. It sounds like, you know, as a buyer, if I have more options and those options are becoming more economically viable, it's just a matter of time before this becomes like ROI accretive. So I think 2024 might be that too. So if these use cases actually start to deliver ROI at scale, what does it mean for talent? Do you actually foresee any workforce implications in '24?

Baris Sarer: Well, that's a great question. The short answer is we'll see limited impact on the workforce but see a new trend emerging. Let me break down these two dynamics. 2024 will be the first year in achieving real productivity improvements as solutions go into production. However, for these productivity increases to translate into large-scale job dislocation, they need to happen at scale and be driven by end-to-end automation.

On both fronts, the market is not ready yet, so net net, I don't expect a workforce impact either in size or in composition of the workforce this year. But an acceleration is likely 2025 onward. Now, AI-enabled automation will start getting factored into hiring decisions. Specifically, we'll start hearing the early versions of the question, "Hey, do we really need to hire for this role? Can we achieve our business objective through automation?" This will be getting asked more frequently at the point of new hiring decisions.

Rohan Gupta: That makes sense. Yeah, so, a very interesting labor market ahead for us in 2024. Maybe no, like, large-scale workforce reductions, but definitely a lot more cautious hiring.

Let's maybe turn to the regulatory landscape, finally, Baris. You know, 2023 began to see some regulation. Where is that headed in 2024?

Baris Sarer: From a legal and compliance standpoint, I predict two axes of battle forming. One will be the legal front, where topics such as fair use, copyright liabilities, etc. are getting hotly debated and yet not fully settled in 2024. On the regulatory front, there are two schools of thought. One is to regulate large language models. The other one is to regulate use cases. For example, the EU's AI Act is an example of the latter. Biden's executive order seems to lean towards the former. Regulating model is inherently more complicated and will be politically charged as it can be weaponized at the international scale. And so, similar to the legal front, these two approaches and their implications will be hotly debated, but I think they will slow roll in effect.

Rohan Gupta: Makes sense. Baris, thank you. These are some incredible predictions. I'm looking forward to 2024, as I'm sure you are as well. We'll definitely have you back on the show in a couple of weeks. Thanks for your time again.

Visit the AI360 Podcast Episode Library to learn more.