Research @ Citi Podcast, Episode 83: AI's Mid-Year Review — The Token Bottleneck

Recorded: July 1, 2026

Published: July 9, 2026

Host: Rob Rowe, Global Head of Research, Citi

Guest: Heath Terry, Head of Technology and Communications Research, Citi

Transcript:

Rob Rowe (00:00)

Today on our podcast, we're going to be addressing a mid-year review of everything AI and innovation. Welcome to the Research @ Citi podcast. I'm your host, Rob Rowe, Interim Head of Research at Citi.

Today, we're really excited and fortunate to have Heath Terry with us, who is our Head of Technology and Communications Research at Citi, and who is absolutely a leader in this field. Heath, thanks for being with us.

Heath Terry (00:28)

Thanks for having me, Rob. Always glad to be here.

Rob Rowe (00:31)

Fantastic. Maybe we can kick in and just look at market performance in terms of AI right now, think about the market in general. I know that there's a lot of questions coming up in regard to how is this transformation going? I think, surprisingly, this market has been on the upside for some time now. It's still going. So, you know, we had a debate sometimes about whether this is a cycle or whether this is just a continuous move in terms of development and innovation. How are you seeing performance in the marketplace right now? And how are you seeing that develop? Is this a cycle, or is this more just a new level that we're going to?

Heath Terry (01:13)

Look, this is a new technology that is fundamentally different from any major technology transition that we've seen before, right? We've been through several of these. We've been through cloud, we've been through mobile, we've been through internet. And this one dwarfs all of those in terms of scale, as we've talked about a lot. The impact that it's going to have on the economy, on corporates, on the competitive stack within industries, is greater than any we've seen before.

And within technology — and this is what you're seeing reflected in the market — this is the biggest shift of economics we've ever seen. We are taking 90% incremental margin dollars out of software and internet and information services and we are moving them into AI models that are right now about 50% incremental margins. And that delta, that 40-percentage-point delta, goes into the pockets of all of the infrastructure providers, all the data-center infrastructure that has to be built for this, which is why you've seen such a dramatic spread in terms of performance between the semi/cap-equipment companies, the hardware companies, the infrastructure providers themselves, and all of the software, internet, information-services companies that are on the other side of it.

So this, I think, goes beyond a cycle because cycles are usually tied to economies or products. This is a fundamental shift in technology, the likes of which we really haven't seen before.

Rob Rowe (02:44)

And Heath, to that point: In terms of adoption, there's been a lot of argument about adoption one way or the other. How are firms adopting this? How is the world adopting this? How are individuals adopting it? Some people would perceive that adoption to take longer than people expect. But do you think that, in fact, it's accelerating? How are businesses looking at adoption? Where is adoption right now? I think that's the other leg that people are also looking for.

Heath Terry (03:11)

Yeah, absolutely. Adoption has been accelerating. We have been saying this for well over a year now, just in terms of the pace of acceleration: It keeps steepening, the bend in the S-curve keeps getting steeper. The best proxy for measuring this, to us, is the hyperscaler backlog numbers. We've gone from 28% year-over-year growth last year to 32%, to 53%, to 91%, to 143% in Q1 of this year.

And that acceleration that you're seeing reflects the adoption of companies at the enterprise level: healthcare companies, financial-services companies, certainly technology companies, industrial companies trying to put this technology to work in a way that is generating a meaningful return on them. And we are seeing that meaningful return show up in higher revenue per employee. We're seeing it in margin expansion across those early adopters. And then, of course, you're seeing it in the returns that the hyperscalers are generating on the investments that they're making into this infrastructure.

Rob Rowe (04:17)

And are there particular industries where you're seeing that acute development happening more rapidly than others?

Heath Terry (04:24)

There are. This is a technology that is very unevenly distributed by industry and within industry. So, there are companies that are further ahead on this than their peers and their competitors. And there are certainly industries.

So if you look at the top in this — and, you know, this isn't going to surprise anyone, but AI has been best for software development: the most performant, the most effective tool. And so no surprise, the companies that have the most software developers are the ones that are benefiting the most from this. So it starts at the top with technology companies, then you get down to financial-services companies, then you have industrial companies where you have software developers and engineers that fall into that same kind of use case.

And then healthcare companies are there, but it's very unfair to healthcare to place it in that fourth-place position, because you have companies within healthcare that are in the world of drug development, or the med-devices side of things, that are very technology-heavy and that are seeing massive benefits from this. And then you've got companies that are more service-heavy, like hospital-services companies, that are elder-care companies that are at a different end of the spectrum in terms of adoption.

And then you get into consumer, where you have big deltas between companies that are on the cutting edge of the adoption of this technology and leveraging it in their logistic systems and leveraging it in their stores with their employees, and then other companies within retail just starting to think about how they're going to deploy and try and catch up.

Rob Rowe (05:54)

I think we had talked previously, Heath, about all the different directions in which AI is providing an application. Can you highlight some of those? So when we talk about a particular company or industry, I know that they're probably looking at ways to be more productive, ways to enhance capacity, but they're also looking at other things. As you said, even in Walmart stores, they may be doing things. What are some of the applications that you're seeing? I know there's some that are physical vs. the traditional way of thinking of this within classical computing.

Heath Terry (06:24)

Look, stage one of this was about cost savings. It was, "How do we make our people more efficient?" How do we leverage our ability to use these tools to automate workloads that used to require people to do that work? Inside a place like a bank, it might be things like tagging credit-card transactions or handling KYC and AML compliance. In a healthcare company, it's managing patient trials, it's managing FDA compliance. Inside of a warehouse, it's better routing for the robotic automation systems that you have in place and the warehouse automation systems that you have in place.

But that was stage one. That was, "How do we reduce costs on this?" Where the leading companies in the space, where the frontier adopters of AI are going, is "How do you use this to grow your business?" How do you use this to accelerate the revenue that you have coming in, to take share from your competitors, to enter markets that maybe weren't commercial for you prior to AI but are commercial to you now, or businesses that weren't commercial that can be commercial now?

And so there's a ton of opportunity here as companies think about, "How do we leverage this technology?" Not to lower our costs — because when you're thinking about it that way, you're automatically dealing with capped returns, your best-case scenario is you eliminate 100% of your costs. Whereas if you're thinking about how you use it for growing your business, your opportunity, your upside in that can go well beyond 100%.

And so, we're in an environment right now, and a lot has been made — I think we'll probably talk about this — about the cost of tokens going up and the cost of AI adoption going up. We're in an environment now where tokens are too scarce to waste on purely capped opportunities. And so if the only way that you're thinking about how you use your tokens is "How do we lower your cost?" you're missing out on the biggest part of the opportunity. And companies are seeing that.

One of the smartest AI developers that I know — and I won't mention him specifically on this, but if you go out to LinkedIn or the web, you can probably find the blog post that he put out there — referred to AI and looking at AI from a labor perspective or labor-reduction perspective as simply companies lacking creativity. This is not about our limitation being "How do we spend tokens?" or "How do we lower our AI costs?" It's about, "How do we maximize our creativity on how to use this?" And I think that's the right answer for the next five years vs. what we've been talking about for the last three years.

Rob Rowe (09:01)

And maybe we can segue then to the cost component, because there's been discussion about token costs and how to reduce them. Are there other infrastructure challenges that potentially impede this progress? Or do you think — for instance, are we building enough data centers? Is the power grid adequate? How are we looking at token costs?

Heath Terry (09:24)

Sure. You know, token costs are a component in this, and I'm as guilty of this as anyone — we sort of oversimplify these things a lot by talking about token costs going up. What we really mean is *output* token prices are going up. Overall token prices, the cost of generating a raw token, continues to decline dramatically. This is sort of the AI Moore's Law that we have, of seeing 90% declines roughly every seven months or so. And as we get the new Blackwell models that are coming on, or the new Blackwell infrastructure that's coming online from NVIDIA — where you've got more efficient token production that's anywhere between 15 to, in some measures, 35x more efficient than the Hopper infrastructure that's out there now — raw token prices will continue to come down.

But because the models themselves are getting so much better — they're getting so much more performant — they're burning more tokens to get to that answer. It's a better answer. And so the cost of intelligence is still going down, but the cost of output token prices are going up. We've gone from $25 on some models all the way up to $125 on some of the models that have been more recently released. That's per million output tokens. You'll see that come down over time as more infrastructure comes on.

But we're in an incredibly infrastructure-constrained environment right now, and that's not going to change anytime soon. We estimate we'll probably build about 15 gigawatts worth of capacity this year in the Western world. We built 11 last year. Our estimate would suggest that we need about 25 incremental built this year. So we've got almost a 10-gigawatt delta between what can be built vs. what's needed. That gap is just going to continue to build because of all the issues that you referenced in terms of power availability, labor availability, electrical-equipment availability.

Rob Rowe (11:11)

And maybe we can shift to regulatory, but also talk a little bit about cybersecurity, because that's another topic that's been coming up with the discussions on Mythos. And how are we looking at cybersecurity? And then is the regulatory environment putting any pressure on this? Or do you think we'll come to some satisfactory conclusions there?

Heath Terry (11:31)

Well, the implication for cybersecurity is really significant here, because these tools in the hands of the wrong people create real risks — and not because they can necessarily do anything that bad actors haven't been able to do in the past. They can do them at scale. You know, there may be ability to spin up the equivalent of a million sophisticated hackers through the use of one of these tools, could just simply overwhelm the cybersecurity systems at any one company. And so that's a real concern on this. It's why you've seen some of the regulatory action taken here.

But, you know, the unfortunate reality is just because you're trying to shut down access to it, or a regulator in one country tries to shut down access to it, the world is not cooperating. These are, in a lot of ways, becoming stateless technologies. And the more we try and regulate and restrict

and control, the faster we encourage that movement towards stateless kind of systems that are out there.

As soon as the world realized that the U.S. government was going to control who had access to these frontier models, the reaction from every country globally was to start investing heavily — if they hadn't already, and most of them had — in their own sovereign models, their own frontier models, their own infrastructure to make sure that they weren't in a position to be cut off by another country that decided they didn't deserve access to artificial intelligence.

Rob Rowe (13:00)

And so, do you see a solution? What would be the answer for that in your own view?

Heath Terry (13:06)

I think it's an easy thing to say, “Well, we're just going to stop everybody from having access to this and that'll solve the problem.” And it's easy to say and it's impossible to do. And all it does is drive more into open source, drive more into foreign models.

And so you have to have kind of a controlled solution. I think what you saw with some of the projects that were initiated to give a small group of companies access to frontier models so that they could begin to shore up their systems and make sure that they were defended before these things got out — the cooperation that you saw between the model AI labs and regulators to make sure that they were coordinated on the rollout here, and that the structurally important industries and utilities and power-grid infrastructure and travel infrastructure, our transportation infrastructure, all of these things were shored up and, as much as reasonably possible, bulletproof before those models got out into the world — I think that is the middle road that we have to come to on this from a regulatory perspective.

And it doesn't mean that we've eliminated all of the risks around this. That's simply going to be impossible. But it does mean that we're at least protecting ourselves as best as we can in the near to medium term.

Rob Rowe (14:20)

And Heath, lastly, one thing I was curious about actually goes back to adoption. It seems like every day new versions are being created, there's even more sophistication being added. For someone who's trying to adopt AI, does that present a challenge, or is it relatively easy on a structural basis to simply update your systems as you go along? So, you know, let's say you adopt one version of AI for one application you have at your firm. Do you then have to worry about the constant innovation that seems to be taking place?

Heath Terry (14:52)

These models are not fully interchangeable. I mean, at a very base level, you can change the API call that's happening on the other end for your agent. But to really leverage AI within an organization, there's a lot of fine-tuning of models. There's a lot of post-training work that needs to be done around internal data that's available. And so there is a kind of component of lock-in that happens on that.

There's been a lot of concern from the beginning about these models being commoditized. We had the same concern about cloud early on, where people looked at what AWS was doing and said, well, eventually you're going to get all these competitors in offering compute and service,

and it's just going to commoditize the cloud and margins are going to fall dramatically because of that. And we obviously never saw that within cloud, or at least to date have not seen that within cloud. And we think that that's probably going to be the case with these models as well, particularly as they move up the stack and offer more agentic services, offer more tools built on top of the models, particularly for medium and smaller enterprises that don't have the capacity to build out a lot of self-built technology the way you'll see among sort of Fortune 500-level companies.

And so, these models, we believe, are going to be pretty sticky, which means they will be profitable. They will have relatively high returns associated with them, even as companies leverage a full spectrum of lower-cost and open-source and even on-prem infrastructure to be able to optimize their AI spend.

Rob Rowe (16:27)

And how important is on-prem right now, the concept of having on-prem vs. not?

Heath Terry (16:34)

Right now it's important in that everyone wishes they had it. But in terms of being a major part of what's going on in AI, it's kind of meaningless, largely because no one has the capacity to build that kind of infrastructure. You've got to rely on the cloud providers. You've got to rely on the neo-cloud providers, because they're the only ones with the scale to be able to service against this. And we're in such a capacity-constrained environment — where there's not enough memory, there's not enough storage, there's not enough interconnect, there's not enough CPUs or GPUs, there's certainly not enough power on all of this — that for individual companies to start trying to build that kind of cutting-edge AI infrastructure at scale on-prem is just going to be incredibly expensive. And because of that, inefficient in terms of their ability to do it.

I don't think you're likely to see on-prem as a meaningful part of this story within the next 12 to 24 months. Longer term, it certainly will be, because companies are going to want to have control, particularly around their own data. And to the extent that regulatory becomes more and more of an overhang about who has access and who doesn't, that's going to require more on-prem infrastructure, because you're going to want to limit your exposure to those governmental choke points. But in the short term, it's just prohibitively expensive to be able to do at scale for companies to really have it as a meaningful part of their strategy.

Rob Rowe (17:59)

Heath, thanks so much for all your insight today. Very insightful. I'm sure our audience appreciates your comments. Thanks again.

Heath Terry (18:06)

Thanks, Rob. Happy to be here.

Rob Rowe (18:08)

This episode of Research @ Citi was recorded on Wednesday, July 1, 2026. I'm your host, Rob Rowe. If you enjoyed our discussion and want to dive deeper into market trends and expert analysis, please like, share, and subscribe on your favorite podcast platform. See you next time!

Disclaimer (18:27)