COLEMAN
VISASQ

Generative AI:
Investable Opportunities and Enterprise Use Cases

Our Expert :
Saad Ansari

Venture Partner at Dash VC

- Director of AI at Jasper Ai, Inc. (08/2022 – 07/2023)
- Director - AI Success at DataRobot, Inc (09/2019 – 05/2022)
- Lecturer at Harris School of Public Policy, University of Chicago (03/2018 – 05/2021)
- Advisor, Obama Administration at US Navy (03/2015 – 01/2017)

Jasper is a leading Generative AI Content Platform. In this role, he led their AI Team, which shipped the AI Engine (modular infrastructure to serve the best model per use case), integrated closed and open-source models, and the experimentation platform - a key to personalize outputs to user preferences. He also started and authored their patent program and pushed the broader product for personalization and end-to-end content management.

Moderator:
Max Le Sieur

Founder & Managing Partner at Rosemont Legacy

- MBA, Harvard Business School
- Investment Banking Associate at BMO Capital Markets (07/2016 – 08/2020)

Expert Insights On:

- Basics of Generative AI: What does generative AI unlock or what can it do that wasn't possible previously? How is it trained? What is the role of a transformer?
- Investable Opportunities: Describing the Generative AI value chain. Describing the AI market landscape and investable areas. What segments have received/accrued the most investment?
- Realities of barrier to entry.
- FAANG positioning and future outlook. Market leaders of the space. Areas that are lagging.
- How do is regulation shaping the landscape?
- Enterprise Use Cases: Which industries or sectors have the most potential? What is the most successful enterprise use case today? Future outlook of use cases and impact.
- What is the biggest impediment to adoption? Including moral, ethical, or legal impediments.
- How should enterprises measure the ROI when it comes to applying AI to their business?
- What are the challenges to training these models? What are the core considerations?

Introduction -

Max: Hey Saad, nice to get connected here. Thank you so much for taking the time. As Alex alluded to, we're really excited to get a bit of your time here. My name is Max and I'll be leading this call on behalf of Coleman. And as you know, the purpose of the discussion is to learn a little bit more about generative AI and to get your perspective on how the space is developing any investment opportunities in the space as well as potential enterprise use cases. Before we begin, I do want to reiterate that we are in no way soliciting any material non-public information or any information that is confidential and related to any company or organization you are currently or have ever been affiliated with.

Saad: Absolutely.

Max: Yeah. If you believe that the answer to any question during our discussion involves any non-public information, please just flag it to me and I'll take us in a different direction right away. And with that being said, do you have any questions for me before we hop in?

Saad: No, I think we're ready to go.

What is generative AI? -

Max: Okay, awesome. So without further ado, Saad, can you please just explain generative AI in simple terms?

Saad: Yeah, absolutely. So generative AI, first of all, artificial intelligence, I define that as an effect of any one of a number of methods including deep learning, machine learning and so on, that produces something that mimics or extends or you could say even provides an alternative to human intelligence. Now what generative AI is, gen AI is that we're specifically talking about the creation of a new form of media. That could be text, it could be image or video or so on, but that's generally what it is.

Max: Got it. What does generative AI unlock or what can it do that wasn't possible previously?

Saad:

Yeah, so before 2017 when there was this famous paper written about large language models called Attention is All You Need. Before then we had natural language generation capabilities, we had natural language process capabilities, but simply put, they weren't very good. What attention allowed us to do in large language models in specifics was to get longer sentence length and even paragraph and longer length that actually made sense and was coherent. And so the main thing is that now longer form text is actually good. It used to be not good. And this was made possible by large language models and transformers. The semantic understanding is much deeper. Also, these models don't require you to train any data. They're ready to go out of the box. They're often pre-trained and they're quite usable. So I'd say to summarize, I think three things.

One is first they're just really great language capabilities. These language capabilities unlock even multimodal capability. You can have very good image generation without good language generation. The second thing is they're just very easy to access because they don't need training data. It used to be that you have to go through this really long intensive, costly process to get your data together, to train some models, to have some sort of value. These models require very little extra training or fine-tuning or augmentation to work. So they're just extraordinarily accessible to power accessibility. And then the third thing, and there's models and there's AI and I think there're actually different conversations in the AI space because we've had a social moment where there's a lot of attention here, we have so many more developer tools, open libraries to make their productization very easy. So I think those are the three things that are different now and it's really exciting.

Max:

Got it. And how does it work exactly? If you had to describe in simple terms like how it works, what does it doing, how is it trained?

Saad:

Absolutely. So there's once again different ways to talk about this. There's the AI and then there's the model. Let's start with the model and move to the AI. So when we're saying generative AI, most people mean the so of the show is the large language model, this model that enables generative AI. But it's not all of generative AI and there's much more to general AI than just that. The way that large language models are distinguished today because we've had language models for a long time, but one of the key ways that they work now is item number one is they're essentially predicting the next word at a time. So you have a sentence, I will go to the... And you have to predict the next word. Large language models, they pretty much all work in that context. How do you predict the next word? They do this using a technique called attention.

Saad:

So before you come up with these probabilities of what's the next word going to be. How do I understand, or in converse, how do I understand what this instruction means? And your coming up with the probability of which word is next. What attention allows you to do very similar to how the human mind works is allows you to know what part of the entire corpus of memory you have to focus on in order to predict that next thing. So attention is like okay, if the context of the surrounding words and tokens is so-and-so, and the next word is this, I can leave out 90% of this rest of the bulk words here and I can make a better prediction of this next word. So attention is a second way how it works. And other than that, there's three capabilities you're trying to optimize for. You could say a semantic capability, memory capability, and instruct following. There's a lot of capabilities LLMs have, but those are probably three of the ones that people are most familiar with.

You could train a model to know all those things, train it to be up-to-date, train it to follow better instructions, train it to be semantically complex in English or in coding or whatever, and have just good English. Moving from model to AI though, just like we always have, you can create systems to, for example, offload the memory into a database and then have the model only focus on semantics and then have the database be responsible for some of the memory. So those are other techniques for the AI. I think we covered just the root basics of it, but this is a topic we could really go on for hours about.

Max:

Yeah, got it. That's super helpful. Can you just elaborate what the role of the transformer is specifically? Because I know that's a key unlock in this whole most recent development of it.

Saad:

Absolutely. So what I just described is what the transformer essentially is. Transformers are often used as what people call the foundational model. So zooming back out into the level of generative AI, a generative AI system for any given application could involve the foundational model which is coming up with a generation. But will that foundational model, that transformer, can it give you your daily financial news digest and analyze what are we good stocks to invest in? No, a transformer alone is a very bad use for that, because it's not trained in that day. That's probably hallucinating. It doesn't know what it's talking about. If you have a foundational model tied with a retrieval layer, like let's just say an API to Bloomberg data, you did some rules-based stuff to parse out the right information you needed. You trained your model to make sure it doesn't hallucinate on stuff from this API when it's phrased in a particular way.

Saad: And then you have some chain of thought reasoning, so that analyzes prices and gives you an analysis of them and it sends you stock prices or whatever. That's more generative AI stuff. It involves more than just the transformer and the foundational model. You don't need the transformer necessarily for memory or even some types of the reasoning, which you can use rules-based stuff or basic math for. But what it is doing is it's playing the role of a logic machine. So here it's generating that newsletter, it's summarizing, but it might also be synthesizing and commenting on, but I feel like the easiest way to describe it is what I described before with a large language model, those tend to be transformer models now. And the role it plays in the AI workflow, it tends to be like the big logic engine, the foundational model.

Investable Opportunities in Generative AI? –

Max: Okay, that's helpful. And then can you walk us through the value chain? So starting with GPUs and then GPUs are purchased by companies that can basically source a ton of data and then design the logic and then use GPUs to train the model. And then from there, the model then can get plugged into additional use cases and then tweaked and improved. Is that the right way to think about it? How would you describe that value chain?

Saad: And so you have the cloud providers, you have your hardware and you have your training environments, you have your model providers, so then you move to inference. So if you're using an open source model, which is let's just say Meta made LLaMA 2, it's an open source model, you download it. You're able to host it on your own managed service for inference, you're able to put it in the cloud. If you're an AI second company, you might want somebody to do that for you. So once again, a cloud can help with that or something like Hugging Face that makes it easy for you. Typically, in your inference engine where you're doing inference, you'll want the ability to build an application layer that is interoperable with that. That's just the general software stack, but you want to make sure that it's interoperable.

If you're using closed source models via OpenAI, you'll just call that API, you'll still have your own extraction there. If you have an open-source model, you'll want to build your own, you could say software infrastructure around that. A lot of companies now as they're getting more and more sophisticated are building things like chaining and model routing and all these sorts of things in the application layer around their inference as well. So anyway, so that's like a bucket of things.

Saad:

Then you have your application there, stuff that you're building that's actually providing value to the customer. And then you have MLOps. Also, I'll say though here I think people over complicate this step, they tend to put the cart before the horse. You only need to do MLOps and LLMOps for that which you actually have deployed to production. I think some people are getting into it, they're getting too heavy into the infrastructure before they have the application usage for it.

So you have some of that, and I'll pause here just the high level. If you're an AI value company, sorry, AI first by value, you might be dealing with things like optimization packages. They make the models have less latency and so on. I can talk more about also how locked in this ecosystem is into the CUDA ecosystem on top of Nvidia. CUDA, CPUs or ASICs and Inference? Happy to talk about all that as well, but I'll pause for now. And also, I'm happy to get into how you train the model with PyTorch and all the things you need for that and feature engineering. But I'll pause for your questions.

Max:

What segments of that value chain have received the most investment? Obviously there's been a lot of hype recently, Nvidia earnings, et cetera. What areas have received or accrued the most value and receive the most investment?

Saad:

Yeah, I mean it's no secret that Nvidia has received a lot of investment. There's a lot of the application, a lot of the model providers have received investment like Anthropic and OpenAI. I'm not very sure about the revenue though, and I think that's why they tend to pair up with these hyperscalers like Anthropic with Google and OpenAI with Microsoft. So they're receiving a lot of investment. If you look at the list of Y Combinator companies, I think a good half are AI software infrastructure like LLMOps, MLOps, and they're receiving some startup investment and I think there are some AI applications receiving investment as well. I think that when you're looking at startups versus incumbents, if you even just look at the stock market now, people have poured into Adobe and Salesforce and not about Salesforce, but like Adobe, Intuit, these other companies that are older and they're going to use generative AI.

I think there's a lot of places where incumbents will stand to win, but I also think that startups have a lot to win, especially in the application there. But to be honest, I haven't seen startup AI enabled applications have, I don't feel like that space is fully taken off yet. People are focused on higher up in the value chain, whereas I actually think a lot of the value is closer to the application there, closer to the end user, but there's probably a couple more months or years left before we get full maturity in that space.

Max: That's interesting. Do you think the application layer that is closer to the user is going to be where the most value is accrued long, long term?

Saad: Oh, absolutely. I mean in the energy sector there's a thing called the smile curve. So if you have your Y axis B margin and then your X axis is the value chain. In the energy sector, it's like a smile. What that means is you have your minerals and mining companies having a good amount of margin, your battery companies making no margin, and then you have your end applications like cars and toys and drones and all that making a lot of margin.

I think it's going to shake out to be the same in the generative AI space. You're going to have your microchips and your real actual hardcore technology providers have a high amount of margin. You're going to have the model providers and some of these, you could say AI infrastructure stuff have relatively low margin. It's getting quite saturated, it's getting expensive, and they're competing with open source as well. So I think the case for them to have a lot of margin is tough, and then you're going to have applicationary companies that are really well-designed having most of the margin at the end. But like I said, I feel like that last part of the smile is ultra immature and we haven't seen good design meet good AI just yet.

Max: Wow, that's interesting. Okay, so that seems a little counterintuitive because it feels like the models themselves and training the model is actually what creates a barrier to entry is actually the difficult part because I mean is that a fair characterization before I go onto the next question?

Saad: It's fair, but it's also temporary. And what I mean by that is, yeah, there was a time when that was the hardest part and that was where all the value was. If you just fast-forward the timeline though, just a little bit, two, three years, you're going to have open source models anybody can use that are more than enough than what you need. It's not like Moore's law applies to these models. You can't just make them instantly better forever because the English language itself has a limit and we're getting close to that. And so you'll have top-notch performant open source models that are very easily usable and we're already touching that moment right now, so commoditized. Also, you don't need to train models to get value, and this is why I make it a point to really distinguish between AI and models.

Saad:

So there's so much more you can do to your AI besides training your model or fine-tuning it to get the value you want out of it, like retrieval augmentation, or you can use an adapter model. They're really easy steps. I could teach a kid to do it in a couple of weeks and you can get a lot of value from that. I'd say the hard part in two years from now, even less, isn't going to be this stuff. It's going to be combining really good design with the end applications.

So for example, I've been advising a lot of startups and I noticed that there's all these, for example, education AI companies coming out, they're great, they have good ideas, but they're all really low hanging fruit like books for kids with AI and this and that for AI. But none of them really think about the design or the pedagogy of how do you teach, what's really possible in generate AI, what are some creative techniques and tools. It's not like Steve Jobs, it's not like the sort of thing Steve Jobs would've made. And so I think that's the hard part, getting good AI tech teams together with really good design people who know the industry and them creating really good products. I see a lot of AI talent. I don't see a lot of design and AI talent.

Max:

Got it. But then how do you differentiate if you're an application that is built on top, if you're just the wrapper for an underlying model that anyone else can use as well. Don't you lose your ability to create something with a moat?

Saad:

Oh, absolutely not. So let's just go back in time to when microchips first came out and we're making computers. Imagine if somebody told an early Apple, hey, you're just like a keyboard attached to a microchip. Shouldn't you be making your own microchips? Isn't that where all the money is? Fast-forward to now, the iPhone is so much more than just a microchip with a screen. There's a lot of complexity in the product. It's the number of microchips working together in different integrated circuits and all that, different sensors and components coming together to make a much better product. And yeah, the microchips are a part of it, there's a couple of them in there, but it's so much more than that too.

Good applications will use a number of models. Interestingly, at the right times, the right places, they'll have really interesting supply chains and customer relationships and all that, and then they'll be doing really cool things. I mean, take Amazon too. Amazon is just e-commerce, it started off there and it moved heavily into supply chain and into turning its infrastructure into something public with AWS and so on and so forth. So no, absolutely not.

Saad:

I feel like the models will be great enablers just like how the microchips enabled iPhone, but I think you're going to get complexity, design depth and a lot of creativity in building these end applications. And I'm not saying these companies won't be able to use their own models and train their own models, but I just feel like that's not where the great innovation is going to come from.

I think there's some areas where it might, but those are not what we're talking about today. It's not generative AI. For example, I happen to believe that you can use models to do microchip floor planning and design better microchips to have better performance and specific uses that are arcane. That's not arguably generative AI. It's just different AI for different use cases altogether. I'm just out of the scope of this question. And the reason I'm thinking that the models will become commoditized is the things that they're doing have upper limits. Your English can only get so good, your images can only become so great, and then everything after that is less the job of the model and it's more of the job of the AI system and the software you're developing using those models or even the hardware too.

Max:

Got it, okay, that's super interesting. Who in your opinion are market leaders today?

Saad:

I mean, nobody's really shocked to hear that Nvidia has a lot to gain here and the cloud providers hyperscalers. Just today, I think those things are indubitably true. I think there's a lot that incumbents do stand to win. So in terms of application, there's Adobe, Intuit, Salesforce just those companies that have a lot of benefit to get from generative AI, they'll do that and it's going to be generally harder for startups to directly compete with them. Because for a startup to compete with an incumbent, the startup has to force the incumbent into an innovator's dilemma. The startup has to do something that incumbent can't or won't do. Whereas for generative AI, it's relatively easy. It's easier for Adobe to integrate generative AI than it's for generative AI company to recreate all of Adobe Suite.

And so OpenAI for sure has the best closed source models in terms of GPT-4 and ChatGPT. Anthropic is a closed second though with the Anthropic models, LLaMA model that Facebook released, sorry, Meta, is fantastic. The Falcon model was really great as well. They have some limitations that the cloud source models don't have, but that's there. But I think the headlines for all of this, the two headlines for this whole space is number one is it's still immature. The industry is still quite immature and 10 years from now it'll feel like we're talking about AOL and Netflix rather than Chrome and whatever the future versions of it are. And then the second thing is I think there's a number of scenarios for how the industry will change.

Saad:

I actually think some of them are normatively better.
Right now, there's a lot of lock into the CUDA ecosystem. I don't think it has to be that way, and I think if it stayed that way that we actually lose out on a lot of innovation because of everyone being locked into a particular way of working. Same thing with the hyperscaler issues, same things with e-commerce. So I think the scenarios where people are creative enough to use generative AI to start real new companies that actually scale and become sizable like the Fan., and I think it's a scenario where people aren't that creative and the incumbents win and we get some startups here and there, but nothing really that changes the terrain.

Max:

Got it. That's really, really helpful. And so if you had to describe the landscape to someone who is an investor by trade but is curious on mapping the environment and the landscape, how would you describe it and what would you say are investible areas?

Saad:

Yeah, I'll start with a thesis and I'll break it down into the taxonomy. So both the internet and generative AI are similar technologies. They're both changes in the relationship between how people access information. The internet changed that relationship from being about a random search through the files, hard files into a soft process, and the main verb, the main action of the internet is index and search. Google is index and search for websites and information. Wikipedia is for encyclopedic stuff. Amazon is for products. Facebook is index and search on people. There's all just index and search, all these giant companies. Netflix is, and it's the recommendation engines are a really powerful thing there. The internet followed a specific timeline where you get the low hanging fruit companies coming out first like Google and you don't need any supply chain for Google, and where the internet is like most of the business and the value proposition. And then you move up and you move up into, you get more complex businesses.

Amazon comes out later, Amazon itself, but it's also a supply chain, so it's internet enabled, but it depends on the supply chain. You get Netflix, which has to do with IP for videos and it uses a lot more compute and it follows this maturity curve where you get these more complex businesses coming out later and the easier ones arguably coming out earlier, taking market share and becoming semi monopolies. Same thing I think will happen more or less with generative AI. You're going to have these infrastructure providers like the microchips, Nvidia, the CUDA enabled ecosystem, some abstraction layer, some tooling. That's already happened, in a way it's a little bit more mature than the application layer. I think it's gotten ahead of the application space because applications usually lead infrastructure rather than other way around.

Saad:

So I think generative AI is a different relationship between people and information. It's a relationship that is based on I think essentially synthesis and routing. So instead of search index and search, there's three flavors I think of effective generative AI. One of them is personalization, another one is co-piloting, and maybe a third one is advanced multimedia like the power of Pixar but for your five-year-old. For personalization specific, what generative AI can do for businesses is, hey, instead of recommending a video, we can create you the best video, we can create for you the best block content out of everything a synthesis. And then routing is we'll give it to you when you need. So imagine a future version of Amazon enabled by generative AI. In this store, you just go and you, hey, I'm going to go to Yellowstone National Park. Maybe it just tells you, hey, I saw you're going to Yellowstone National Park.

It knows you, so it's able to generate things that appeal to you. Here's a future itinerary for your perfect trip to Yellowstone. You have three kids, here's things you can do with kids, blah, blah, blah. You speak this language, here's a really good restaurant that people that speak your languages like or whatever. And by the way, you should buy these waterproof shoes or you should get this coat, because it's winter and you live in Florida and you don't have any coats. And it basically it's learned that you are more likely to read that material that you are more likely to buy when we recommend one product for you in this place here. And it has a 40% higher sell rate than Amazon, and the inference is really cheap and it sells to you. So that's an example of synthesis and routing. It gave you the thing you needed and it synthesized information and it had a really great objective uplift in its business metrics of higher sales and conversion and so on.

So top line, the investible areas of generative AI, I would look at how the internet progressed, look at every single business the internet produced and which ones were big, which ones were small, and just anticipate basically the same thing happening again. There's probably going to be a Facebook competitor that uses generative AI to bring people together. Probably going to be an Amazon competitor, probably going to be some sort of a new form of YouTube where the videos are maybe a little bit more interactive and generative for you. Stuff like that.

I think that's where there's going to be most of the... You look at the complete and total pie that a value created, I think a good chunk of it's going to be in these applications that are FAANG competitive, and then you're going to obviously have the microchips and CUDA layer being another area to look into, but only if the Nvidia, CUDA ecosystem is interrupted a little bit, it'll open up more businesses there. And then you have the model providers and these LLMOps tools .

Saad: Honestly, I think they're great and they're essential, but I also feel like they're going to get harder and harder to monetize. So that's my general perspective on how it might shake out.

Max: Just to push you a little bit on this, are we sure, is the thesis that there's going to be an alternative to the FAANGs that are entirely digitally generative AI native? Or is it possible that the FAANG companies are successful in... Or is the thesis that the FAANG companies to your earlier point about Adobe, are better best positioned to add generative AI as opposed to having the innovator's dilemma play out where a generative AI company out competes them?

Saad: I think there'll be a bid to dethrone all of FAANG. I think it'll be case by case. Like I said, I don't really think in absolutes, I think in scenarios. I think the scenarios where the FAANGs are vulnerable are when, A, they're able to be forced into an innovator's dilemma. Meaning for example, will Amazon change its entire storefront to be more of an experience based selling platform? I don't think so. So I think the e-commerce version has a fighting chance to make some niche. But other things too, execution risk. What do these people are good at generative AI, but they don't care about supply chain, that's like a recipe for failure.

So I don't know if they're going to succeed or not. And like I said, there's scenarios where the incumbents win. I think that's probably the main scenario that the incumbents end up winning. I think that there's scenarios where the startups win, but it takes an intense amount of ingenuity and so on. Unlike other innovations too, the companies that were born with the internet age are a little bit better poised than their predecessors to adopt this new wave of technology, generative AI and so on. And they've already shown an intense willingness to do so. So there'll definitely be bids to take them on. I don't know if they'll succeed or not. That just depends on so many different factors. There's also going to be new use cases that are possible that don't have to do this, I happen to think genomics and AI is poised for some really interesting activity as well, or healthcare, we don't have a [inaudible 00:33:39] so some of those things might happen as well.

Max: Yeah, that's good. Okay. That was super, super informative. Are there areas with regards to the infrastructure or any foundational technologies that are lagging in order for generative AI to be more broadly adopted in your opinion?

Saad:

I think the main thing that's lagging is what I call the demand generation. I advise a bunch of companies that are actually really good at some of the infrastructure and hardware and they have some amazing proof of concepts. The problem is that there's no application that can use them. So for example, take my future Amazon example where you have millions of customers going onto this website or app or whatever, and it's providing them exactly the information that they wanted and the way that they wanted perfectly tailored for them to love reading it. It's like their personal Atlantic or something, some great newsletter just for them. And it just sells them stuff and it happens to sell them stuff at a really great clip and the margins are high and all that. So if that happened, essentially this business would be doing insane amounts of inference. Every time somebody goes to a page, that page is created for the first time ever for that person. Maybe it stores a couple of things and it has some really interesting data retrieval mechanisms that make that cheaper. But essentially you're multiplying the amount of inference you do.

I mean the copilot use cases, so create my marketing copy, create this, create that, create my code. Those are limited, because you're still limited by output of what humans can do. This is like readers. You're creating for readers. There's like thousands X more inference and it's happening automatically and dynamically. What you would need from an infrastructure perspective for that is the right optimized experimentation techniques to make sure the outcomes are actually good and measured and you're constantly creating the best outcomes and prompting the best outcomes for the generation. You have maybe microchips that are more efficient and faster. You have the models that are fit for those sorts of microchips. It requires all sorts of infrastructure. I don't think that's what the problem is. I don't think it's like people want to build this and the markets just waiting for them and they're just waiting for their infrastructure to catch up.

I think it's the other way around. I think a lot of the ideas for the infrastructure exists. I think there's people who have [inaudible 00:36:03] and they're just waiting for people to use it that way. Like I said before, I think the biggest dearth I see is from the creativity in the end applications to then demand this infrastructure. There's no big scaling personalized Amazon in generative AI right now. They're still getting their footing, and so I think the infrastructure is waiting in the wings, and then I think once the demand picks up, then you'll see the infrastructure scale with it.

Max:

Yeah, that's good. Okay. That was super, super informative. Are there areas with regards to the infrastructure or any foundational technologies that are lagging in order for generative AI to be more broadly adopted in your opinion?

Saad:

I think the main thing that's lagging is what I call the demand generation. I advise a bunch of companies that are actually really good at some of the infrastructure and hardware and they have some amazing proof of concepts. The problem is that there's no application that can use them. So for example, take my future Amazon example where you have millions of customers going onto this website or app or whatever, and it's providing them exactly the information that they wanted and the way that they wanted perfectly tailored for them to love reading it. It's like their personal Atlantic or something, some great newsletter just for them. And it just sells them stuff and it happens to sell them stuff at a really great clip and the margins are high and all that. So if that happened, essentially this business would be doing insane amounts of inference. Every time somebody goes to a page, that page is created for the first time ever for that person. Maybe it stores a couple of things and it has some really interesting data retrieval mechanisms that make that cheaper. But essentially you're multiplying the amount of inference you do.

I mean the copilot use cases, so create my marketing copy, create this, create that, create my code. Those are limited, because you're still limited by output of what humans can do. This is like readers. You're creating for readers. There's like thousands X more inference and it's happening automatically and dynamically. What you would need from an infrastructure perspective for that is the right optimized experimentation techniques to make sure the outcomes are actually good and measured and you're constantly creating the best outcomes and prompting the best outcomes for the generation. You have maybe microchips that are more efficient and faster. You have the models that are fit for those sorts of microchips. It requires all sorts of infrastructure. I don't think that's what the problem is. I don't think it's like people want to build this and the markets just waiting for them and they're just waiting for their infrastructure to catch up.

I think it's the other way around. I think a lot of the ideas for the infrastructure exists. I think there's people who have [inaudible 00:36:03] and they're just waiting for people to use it that way. Like I said before, I think the biggest dearth I see is from the creativity in the end applications to then demand this infrastructure. There's no big scaling personalized Amazon in generative AI right now. They're still getting their footing, and so I think the infrastructure is waiting in the wings, and then I think once the demand picks up, then you'll see the infrastructure scale with it.

Max:

Yeah, got it. How do you see regulation shaping the landscape?

Saad:

Yeah, I recently went to a NATO event and I gave a talk about European regulations and so on. And one thing that was notable is you're getting these different flavors between US, Europe, and China, and definitely the regulation is going to inform the flavor of AI. I think it's generally safe to say that China's going to go for a more authoritarian version of AI already heavily the use for surveillance and so on, and that's where some of the investments going to go. I think Europe and America are still getting smart about it. I think it's great that they're thinking about data privacy. I noticed that some of the regulations, they're not outcome focused. Like, hey, did we actually prevent X, Y, Z sorts of abuses? It's more of like it's trying to inform the how, the exact methods, and I don't view that as being effective.

It's pretty hard to inform the outcome through the input method. There's always ways around that. And then they might have unintended negative effects too in development. So definitely there's a lot of power there to affect where it goes. My hope is that Europe and the US, they lean into their main strength, which is strong civil society, strong people. And instead of having fear-based regulation, they have more of the best defense is the best offense and the best offense is education, and making sure we invest in education and STEM so kids are comfortable with generative AI. I think that's the thing to look for, that they invest in open source. In Germany, for example, right now, professors who innovate in AI, they can't start their own businesses. So regulation like that impedes innovation from happening there.

So I still think it could obviously heavily characterize how the AI looks in each country and then even what international affairs looks like around these topics, and it's really important now, but I feel like the US and Europe are still getting their heads around it and tech people and policy people really do speak past each other. I've been on both sides. I've been in the Obama administration and I've been AI lead, and when I listen to policy people speak, I'm like, oh, they don't really get the tech. When I listen to tech people speak, I'm like, oh, they don't really get the policy. And it'd be better to get them both together and hashing them out for better outcomes, outcomes based on hope rather than fear.

How can Generative AI be applied to Enterprise use cases? –

Max:

Awesome. That was super insightful. I want to turn our attention a little bit to enterprise use cases because it sounds at least today like these enterprise use cases are going to be one of the killer productivity enhancements to the economy generally if when AI can be applied into all of these different pockets of the economy. And so I'm curious what you think are the industries or sectors where AI has the most potential?

Saad:

Yeah. Let's see. Back when I was a DataRobot, I served almost every industry with their AI journey and generative AI does definitely change some of that. I'll start with the one vertical I have the most hope in. I think healthcare has been relatively resilient to innovation and genomics as well is super interesting, but its promise isn't fully unlocked yet, and these things are generally unproven. I think there's a possibility for healthcare in general to be reformed. There's a really good book about it by Eric Topol, I think his name is. Yeah, Eric Topol. Healthcare is not about AI, it's about health and care. It's about people and it's about empathy and all sort of stuff. That said, there's ways to, if you're thinking about it holistically, you're thinking about the entire healthcare value chain diagnosis and from diagnosis, patient care prevention and all that, and discovering treatments to rare diseases.

If you have a holistic approach towards that and you're using generative AI as a part of that more holistic approach, which involves so many more things in generative AI. Maybe it could help convene and think about broader change so that we're able to diagnose earlier. We're able to address rare diseases. We are able to have more preventative care interventions early on. I mentioned this because there is a really interesting intersection with genomics and AI and diagnosis and preventative care, but the value is not fully realized unless other things change too. Just because the AI part happened doesn't mean the rest of the change management has happened. Moving into other verticals too, at DataRobot, I saw that there was so much, and I'm sure a lot of people on the call even as well, or people who are listening or anyone listening, old AI, people had a lot of promise around things like predictive maintenance and your fleets will be perfect and supply chain, and anti fraud, these are the older use cases too.

They're still valuable, but what they teach you is that even when AI succeeds at doing what it does, the prediction degeneration, whatever, that you still have often a lot of last mile change management left, that's often the real blocker. Just because you predicted you needed three parts in this state doesn't mean you have the engineer to actually go change those parts or that the parts that are available. And so those things haven't changed.

Saad: Now, to answer it from the flip side of what industries besides healthcare could be really revolutionized by this. So I would like for healthcare to be revolutionized. Other than that, you've already seen the low hanging fruit start to get knocked out, like marketing personalization around marketing. Going back to my three types of generative AI outcomes, one is co-piloting. Code co-piloting, AutoCAD architecture co-piloting, tax co-piloting, you can co-pilot anything. And what this does is it not only makes people more efficient, but it lowers the bar of entry.

So now you can have non marketers get into marketing, you can have non coders get into coding, you can get non infrastructure people get into infrastructure, and that's really the benefit. It's not that somebody did their job 30% faster, it's that now you can collaborate with more people and get the designers to help you code and maybe you'll have some different products. And so that's the bucket, the Pixar bucket where you're coming, the power of Pixar for your five-year-old. I do think that's quite possible, and I think 3D objects have an advantage over other types of things because for multimedia, you need essentially control and output, meaning you need to be able to manipulate what you're doing, have character consistency and so on and so forth. So 3D objects I feel like are probably going to mesh really well with generative AI and you're going to the Pixar type thing and media will be changed. And then the whole personalization thing, so e-commerce and books and supply chain type stuff. So all that will come with time.

Max: What's the most successful enterprise use case today for you?

Saad: The most successful enterprise use case today, I mean, I used to be at Jasper and Jasper had a huge amount of revenue success with generative AI early, and there's a bunch of copycats now, literally copycats because it's about marketing copy. So just from the pure marketing has been successful, it's the low hanging fruit for generative AI. It meshes really, really, really well. But like I said, the market's immature and I hope a couple of years from now, there'll be so many things that will shadow that. Because I love marketing, but healthcare and education supply chain, these other verticals are just, I think objectively more valuable to human society.

Max: Got it.

Saad: FinTech.

Max: Yeah. So in your opinion, the real success for AI is healthcare.

Saad: Imagine what other use case can you say, this saved millions of lives. This saved you XX percentage of your family's entire lifetime projected income. The only use case that can have such a radical positive effect is healthcare and it's possible.

Max: Got it. That's super helpful. If you thought about the future in 10 years, if you were teleported to the future in 10 years, what would be different if you were right, you know what I mean? How would you be able to tell that, oh yeah, people clearly figured out how to apply AI to healthcare?

Saad: I think Walter Isaacson will be writing a biography about the Steve Jobs regenerative AI, and I would love for that person to have been in healthcare where it's like, oh my God, look, this person, this group of people, they've saved millions of lives. They use this research with AI enabled genomics and this early diagnostic thing and then this coaching thing and so on. They made an entire new healthcare value chain that focuses on prevention, focuses on rare diseases, and it focuses on family-based interventions that were unlocked through genomic side effects. And it was just a mastery of design and hardware design, software design process design and all that. This was Steve Jobs level accomplishment. I think that would be one really wonderful sign. I think also if you got FAANG competitors or just if anybody can access anything, so your five-year-old can make a Pixar level video, all those tools are accessible and you get these competitors to FAANGs, I think it's another two indicators.

Max: Got it. That's super interesting.

Saad: But the main one really isn't about the money or the market caps, honestly, because you can have companies that won, that does mean that they were the best possible company. For example, today, the biggest phone company could have just been terribly designed. It just happened to have been the biggest phone company. That's not a success. I feel like combining the highest level of design possible with the application, that's how we feel about the iPhone typically, or anything that comes out of Apple. That's really, I think the thing to look for. Was it the best possible company that monetized and became huge and was it the best possible design? I think that's different from just the big market cap that became the next Google or something.

Max: Makes sense. You mentioned something earlier that I thought was interesting where it's like the last mile change is still the biggest impediment for the real widespread adoption or for the real reliance on AI within an enterprise use case. Would you say that that's the biggest impediment today? It still only gets you 80% of the way there.

Saad: Sorry, by last mile, do you mean preventing hallucination and the AI value production? Or do you mean what I said earlier about demand generation?

Max: I guess the question is, what do you think is the biggest impediment or the biggest problem or friction that people still experience? And is it that last mile piece that you described earlier whereby it still only gets you 80% of the way there, or is it something different?

Saad: Yeah, the biggest impediment, so in terms of the specific problem with AI hallucination, getting it to know the knowledge you have and all sort of stuff, I actually don't think that's the problem in terms of great growth for enterprise value, because I think just from an engineering perspective, those things are solvable. You might have to scope down your use case a little bit, but that doesn't mean you couldn't get crazy value right now. So I actually think those are relatively solved. I think the big problem has to do with talent and the way that we think about AI talent. I see so many teams that have fantastic engineering skills and AI skills, and they're just super-duper smart. But what I don't see a lot is creativity around applications. There's some that are really good. I like character.ai. They were creative in how they made their chat box and they have some good design into it and so on and so forth.

But it's rare to find a team that's design savvy and they know their users and use case and they have some really deep insights into the market. Mixing that with the tech talent and the AI talent, that's rare. Even at places like Google and other places, it is actually surprising how you won't find that integrated talent even in some of the bigger companies anymore. And so they're in different worlds and when those worlds come together is when you get the best product. And so I look at a lot of teams, startup teams. I've seen a lot of bigger teams, and it's just one of the things that's rare, when you find it's amazing and you love it. And those teams do end up getting a lot of attention eventually, but it seems like we have a lot of tech-centric teams that are hammers looking for nails, but not those things coming together. And when those things familiar, you get demand generation, the rest of the space takes off too.

Max: Got it. Got it. What do you think are the biggest moral, ethical, or legal impediments to enterprise adoption?

Saad: Definitely consumers care a lot about privacy and they care about it, not just its reality, but also it's perception. So for example, a lot of consumers will say, I do not want my data to go to Microsoft or a specific company or like, hey, if my data goes here and I don't have this protection, will the model be trained on it or will the model remember it? That's actually just all perceptional because even OpenAI has a policy that the data is not saved, it's deleted, especially if you opted into that. And then models don't learn your knowledge when they do inference on your knowledge unless you design a system that allows for that. But they still want these guarantees. They still want to have their own model for that reason. So it's safe.

So there's a lot of privacy perception and request based on that, and that's definitely shaping how companies address them. I've been a little bit saddened when I see companies run with that without educating the customer. It's fine to do that for the customer, but you usually tell them, hey, that's not how it works. There's companies that are monetizing just on the privacy fear, and I don't think that's right. It's not a good thing to do for the consumer. And so that's definitely a big concern that consumers have.

In terms of ethics, definitely a lot of authors are getting very concerned about their data being used to train the models. I do think in a more mature space that authors whose information or data or whatever was used in the model, they should be able to get royalties and indeed turn that bug into a feature. Why don't people create specifically for AI to evoke their writing style? And that's a monetization opportunity. I think that's great. But right now it's more of a bug than a feature. It's like people have legal issues about that rightly so. And in terms of other moral ethical things, there'll be effects for jobs for sure. But like I said, I think the right posture for that is a hopeful one rather than regulation. Education is more so the answer there. And this is so easy to get into this space now too. We're talking about a couple of weeks of training rather than years and months of a PhD. So I think the answer there is hope rather than fear, because I don't know if the fear based approach will work out for us in the long run.

Max: Got it. Super helpful.

Saad: Does that answer the question?

Max: Got it. Fair enough. Going back to, I have a couple of questions here on how they're trained. So we're going to take a step back a little bit and just double click on the training. Can you explain just how it works and what the challenges are with regards to training these models?

Saad: Yeah, so let's see. Let's say there's Training proper and then there's training lowercase T. So Training proper is actually the process by which a model, you have your AI engineer going in there using a framework like PyTorch. They've actually identified training data, they're creating a model from scratch. They're going through a number of approaches to create this base model. They're manipulating the weights, they're cycling through different data sets to get an outcome that they've predetermined through scientific methods. It's like that. And then I think a lot of people say training to mean the AI doing something different. So I won't talk about that now. What I'll say about training is, one, most companies don't need to be doing it, especially AI users. For any sort of outcome difference you want more up-to-date data, specific tone and voice for you, all these sorts of things, there's usually an easier way to get it than route training or even fine-tuning.

I think fine-tuning can be very useful for that, but there's ways to approach success even before you get to fine-tuning. But anyway, so just talking about training proper, like that process, the two easiest answers are, one is access to compute, which is costly, and the other one is access to the right data, which is also hard and costly. And maybe the third one is actually access to the talent that's able to do it well, which is also expensive and costly. So those are the three challenges with training. Does that answer the question or should I also address the second piece, which is, most people, I think when they mean training, they mean customizing the outputs for their business.

Max: I see. That's an interesting distinction. Can you elaborate? The three things you said for Training, big T Training are talent, availability, data, and then what's the third one?

Saad: Compute.

Max: Compute, got it. And then can you double... Sorry, go ahead.

Saad: All three boiling down to a lot of money basically, but it's getting cheaper, but it's still pretty expensive.

Max: But the data piece is the one thing that's differentiated and not easy to do in theory, right?

Saad: I mean, the talent piece is pretty hard too. But in terms of the data piece, yeah, it is differentiated. But also as we go forward in time, the value of core root training goes down and the value of the last mile stuff goes up as you get better and better pre-trained based models to work with. So that's true. Right now the data is a big issue, but almost already, but also more in the future, you'll need less and less data in order to do the specific customizations you need for the model.

The only reason that a big company R&D center would want to train a model again from the root or fine tune it heavily to get some other things, what I call a capability difference, higher semantics, higher token contact window, better instruct following, but more different logical tasks, so on and so forth. Most smaller companies or even a lot of Fortune 500 companies, they shouldn't really be messing with that stuff because they're not likely to succeed. And even if they do by the time they succeed, somebody else would've done it for them. And so that's why I make a distinction here. I can elaborate the second half if I answered the first part correctly though.

Max: No, that's super helpful. Please go ahead and elaborate on the second half.

Saad: So I've heard when a lot of users use the word training, what they're saying is, how do I get my tone style knowledge base, intended behavior, so on and so forth into my AI? Going back to my previous definition, a model is a model. AI can be a lot of things. AI can be a foundational model that ties to a retrieval model, that ties to an adapter model as well, or specific prompting or routing. If people just expand their definition a little bit what goes under the hood, they get a lot more optionality. So before you get to training, maybe even before you get to fine-tuning, you could use prompting training, you could use retrieval method to probably get to maybe, I don't know, 70% if not a hundred percent of where you need to go for your outcome for end use case. And not only is it cheaper and easier, but it's also highly, drastically more maintainable.

So for example, imagine a world where you're fine-tuning a model to know the latest financial information. You need to retrain that thing all the time, every day, and that's expensive. And also it still hallucinates, and it would be an unwieldy process because little errors could drastically change the quality of the outcome.

Saad: Whereas if you have a model, maybe it was fine-tuned once or maybe it's already really good instruction following where you're like, hey, I'm going to give you this API's information that's prompting. I'm going to make a retrieval layer and do blah blah blah things to make the retrieval layer work. Then every time you do inference on that, you'll get the right outcome that you wanted for your newsletter scoped on the use case. But then also you have to do nothing to update it because the API automatically updates the model anyway.

You can do the same thing with adaption where, hey, I could retrain something completely in German or I could just have the output be in English. And there's a small secondary model that then translates that at the end. You obviously have to look at the trade-offs, higher latency and costs and this and that, but nine times out of 10 there's an easier way to get a goal done and a better way objectively than training a model or fine-tuning. And like I said, Fortune 500 companies and small SMBs, if they're actually training a base model themselves, that's a large language model. The other base models you can train for sure, like time series models or something, but it's a really high bar to jump over to justify that I feel like in a lot of cases.

Max: Yeah, that makes a ton of sense. What about, is there a quality over quantity or, sorry, going back to big T Training, on the data piece, is there a quality over quantity consideration or a thing that matters there or not really?

Saad: Oh, absolutely. I mean forever I think people focus on big data rather than the right data, and so the process of figuring out what data to get and how to get it and how to structure it before you train it to affect your training outcomes, it is a field called feature engineering, which I love. I think it's a really wonderful field, but I feel like, I don't know, I feel that people don't really get into that field too much. I think, I don't know, maybe it's considered boring, but the whole point of good feature engineering is such that the data you picked has the effect you wanted, but better semantic quality or so on and so forth. Where we are right now in training root models, in terms of our sophistication level, it started off as what I call flame throwing a cow because you wanted a steak, instead of cutting out exactly the piece of the cow you wanted to cook and cooked exactly how you wanted it, we flame throwing a cow.

That's how the first large language models were made just because we didn't have much sophistication. We're just jamming things down and we're learning new methods, which it's not a mistake that we had to do that to get where we are now.

Saad: But it was flame throwing a cow at the end of the day and that was the big data and the massive data approaching techniques. We're getting more and more sophisticated both in terms of the model training process and our sophistication with that. But I think as well as with feature engineering, which is the right data selection and getting it down to a science, I think we still have a long way to go before feature engineering for LLM has become a complete science and it's relatively predictable because there's a couple moving parts to that. But I think people are becoming aware that that's where it should go and then right data becomes more important than big data.

Max: Got it. Well, Saad, thank you so, so much for your time today. That sums up questions up had. Again, we are extremely thankful to get an hour of your time. I mean, I can't think of a better person with whom to have this conversation. This was fantastic and really packed with insights, so thank you so, so much for taking the time.

Saad: Absolutely. It was a total privilege. I really enjoyed the questions and hope you all have a great rest of your day.

Max: Awesome, Saad, so we'll be in touch, but thank you again.

Saad: Thanks.

Max: Cheers. Bye.

Saad: Bye.

This Transcript is accompanied by Coleman Research's comprehensive attestation completed by the Expert following the Hosted Event conference call (the "Attestation"). The Attestation requires the Expert to re - confirm, inter alia, their qualification to consult with CRG in accordance with: 1) Coleman Research's Expert Terms & Conditions, 2) any duties, agreements or contracts in connection the expert's employment, or otherwise, 3) the absence of any disqualifying events in the Expert's personal or professional life, 4) Coleman Research's Seminars restriction against employment by or prohibited relationships with any company with publicly traded securities or government entities. Finally, the Attestation requires the Expert to re-confirm that they did not discuss any information of a confidential nature or provide information constituting material non-public information as circumscribed by applicable securities laws.