accenture

EP. 55: PRODUCTIVITY TO CREATIVITY, THE REAL VALUE OF GEN AI PT.2

AUDIO TRANSCRIPT

[00:00:00] May Habib Enterprise ready Generative AI as secure, accurate, legal and transparent.

[00:00:16] Theresa Tung Hi, welcome to another AI Leaders podcast. My name is Theresa Tung. I am Accenture's chief technologist for Cloud Data and AI. This is part two of a two-part session where we're looking at Accenture's case study about where we're applying Generative AI in Marketing and Comms. Our first part episode was about is the value real. And in the second episode, we're going to look at Can we trust generative AI? I'm so thrilled to be joined in the session by May Habib, the CEO of Writer, and with Marc Appel, who is pioneer in how we apply generative AI in Marketing and Comms. I want to set the stage that Accenture, we're not only a user of Writer, we're also an investor. So just this September, we announced our investment in Writer because we believe Writer offers this easy button for our clients to jump GenAI. And so, May could you start with sharing some more details about how Writer makes GenAI enterprise ready.

[00:01:18] May Habib Enterprise ready Generative AI has got a number of attributes. We see enterprise ready Generative AI as secure, accurate, legal and transparent and what we mean by secure is the data that your users are going to be putting into the Generative AI platform, not going to leave your environment. If it is going to leave your environment into a Writer environment, have you looked at what use cases you're going to allow to be managed in a multi-tenant or single tenant third party environment versus you need to stay inside of your virtual private cloud? We allow for both, and you can think about use cases you built on Writer are almost like headless AI, depending on the nature of the use case, the data it needs to access, etcetera, you can decide which environment you want those use cases in. But the security of the solution, the security of the data that is used to fine tune the models, to train up the use cases that the models access for generative purposes that your users are putting in, all of that needs to be secured and compliant within your privacy and regulatory framework. That's number one. Number two, this is all table stakes, is... Is it accurate? Like, is the generation high quality so that your experts, your professionals say, damn, that is good, that might even be better than what I would do. Because if it is not that good, then they are not going to use it and it's not worthy of their time. It's not worthy of your change management effort. Accurate also means that if I am asking the LLM product I built a question and there is like one answer to that question, it gets it right. And enterprise data environments are messy. It's a lot of unstructured data, there's a lot of repetitive data, data gets outdated very quickly and the type of solutions that you build have to take all of that into account, and that's number two. Number three, and I think we're going to hear a lot more about this as soon as the EU AI Act becomes law, the training data that your generative AI products use is going to get a lot more scrutiny. And with the EU AI Act,

with the fact that fair use for training data in Europe looks a lot different than it does in the U.S. and other countries, compliance with regulatory frameworks internationally is just going to be... It's going to become really important, and I think people are going to be asking much deeper questions around legality and regulatory compliance in the coming months and years and that's got to be part of enterprise grade generative AI. And lastly, is it transparent? And this has got a number of meanings. Generative AI is the only field right now where the vendors and the customers are co- building the future of this technology. The capabilities of these models are accelerating, which means what the models can do, what the technology can do is outstripping our capabilities of like, our abilities to figure out what to do with the technology. And so, we are partners in so many ways and that mutual transparency is so important, whether it's working with a Writer, working with an Accenture like these three teams, working around the client, just have to be incredibly transparent with each other to build useful, powerful solutions that actually get business outcomes. And so, transparency has got to be at every level and from a technology perspective, you know, we see enterprise grade generative AI as transparent with regards to code, with regards to training data, with regards to model weights even. And it is really hard, I think, to provide that level of transparency if you don't really understand the needs of the enterprise and so those are the four requirements: secure, accurate, legal, transparent. From a product perspective, there's a whole set of requirements, we think, around auditability and explain ability, collaboration, being able to really memorialize best practices for a team. There are a lot of requirements we have written those out for folks who are curious in a checklist, it can be very overwhelming, but if people remember secure, accurate, legal, transparent, you know, they're 70% of the way there.

[**00:06:33] Theresa Tung** So insightful, May, and, you know, this is something that Writer does, but it's something that is a good check list for anybody looking to either build their own generative AI applications and models. Same for us as we look to work with our partners, we should be asking all of our partners the same sort of questions, right? With Marc, right, responsible AI it's been an integral part of Accenture's practices and even embedded in our code of business ethics since 2017. Now, with generative AI, as you were looking to adopt generative AI in Marketing and Comms, what are some of the things that you did to be sure that we were responsibly implementing generative AI when we're deploying it at our enterprise scale?

[00:07:24] Marc Appel Yeah, that's a great question. I think, you know, when we first started looking at the type of use cases that we were trying to achieve, the good news as we, again, as we started first with, you know, the productivity, low hanging fruit that everybody can bite off. You know, a lot of the use cases that we were using were not...we weren't as worried about confidential data because we were focused on sort of generating net new in a lot of cases. But where it quickly got to as we expanded outside of use cases was, wait a minute, this needs to be our own stuff, right? And we needed to be accurate, and we needed to be approved and we needed to be derivative of something that we ourselves are authoring as sort of as a group. So from a, you know, both the internal security and internal data structure of what we were going to use, we wanted to make sure that it was well formed, that it was put into...we sort of had some principles around what type of stuff we were going to start with before we actually rolled out in a bigger way because we were worried about - and, you know, this isn't just in Marketing and Communications and I don't, just to be clear, I don't sit across the other functions - but within Marketing, you know, we kind of touch everything, right? We're working with investor relations, we're working with, you know, all the Corp Comm stuff, the 10K. We help our CEO with her earnings communications so there's a lot of - the cybersecurity group, as you can imagine of a big practice, right? We can't let anything leak out or it's bad in a lot of ways for us. So, as we thought about the use cases, we started with the ones we knew we were going to be low risk, right? That had high volume deliverables, that were low risk, that we had the right type of data for. As we started to get that stood up, we then pivoted and said, okay, let's get the process set up for those use cases which we know are a little bit higher risk. We made

sure to engage our Center of Excellence to evaluate those use cases ahead of time so that they could flag whether the data, not the integrity parts, the security part, would be a problem from an external audience standpoint. And then we looked at, okay, from an integrity standpoint, how are we managing? You know, there are 14 different ways to say the definition of one of our key themes, right? So, who's guarding that data? Right? Who's the data librarian that is looking at this stuff and going, yes that, not that. And so, to be completely candid, we're still working through that right now, but we've got a pretty good handle on it between a few different groups that we've now assembled to ensure that when we train our data, we're training it with what good looks like. And I think that's the critical thing, is establishing that organization within what you're trying to do that will govern it. If you don't have that, I can't even imagine what you're going to get out the other side of it.

[00:10:26] Theresa Tung Well, it's clear that Writer also takes AI guardrails very seriously. May, can you share some of the additional guardrails and certifications that you see.

[00:10:36] May Habib You know; the certifications point is interesting. There are no generative AI certifications right now. There is a New York City Bias Act, which we are actually getting audited for, and there will be the EU AI Act but so far, no kind of regulatory sort of certification bodies have really kind of risen up. And we've looked at, you know, there are a couple folks that, you know, talk about AI certification, etc., and we're like, sign us up, examine us, you know, like we're all looking for ways to make it less confusing for enterprise buyers to understand. But really, those tools are just a way to see, you know, like what open source, what products use which open-source models, like its honestly still kindergarten stuff compared to what enterprise teams are really diligence, enterprise security teams. And I do think there is a range of kind of generative AI preparedness on these security teams from folks who, you know, are leaned in to support the business, some of the business applications have taken hold already, and they really just want to support those folks doing all that securely and experimenting securely. You know, to the other side of people who are really scared that the crown jewels will leak into some model and everybody around them, including themselves, will get canned and are saying no to everything. And I'm excited for 24 to be, I think, a year of security innovation around AI, and it's going to benefit Writer, it's going to benefit everybody who's trying to make decisions. But people have got to get their act together because like I haven't started to really see a movement on this front. I would love to see like a SOC 2 equivalent for generative AI, where really, the requirements to be secure are well understood and that gets reflected. So, in the meantime, what are we doing to sort of, you know, hold our hands and hold ourselves up to this really high standard, you know. I think first there is a ton of risk around prompt jocking, data exfiltration through prompts, and those are real risks. And we do things that we don't think any other LLM does around screening. What comes out of a Large Language Model for outgoing links and harmful stuff that we think, you know, a user might get tricked into doing or clicking, etc., that might, you know, kind of tricking LLM into doing something. So, like there is a whole set of questions I wish we were getting asked on stuff like that that we do. And so that is part of our AI guardrails. Part of AI guardrails is being compliant from a content perspective to a lot of regulatory requirements based on industry. So, in, for example, financial services, our customers can't talk about things being free or things being guaranteed and there's like real specific language that has to be used and if you're using generative AI and doing things at scale, while you may not necessarily be doing things slowly enough to catch this stuff, so that all has to get built in and you can't really just kind of ask nicely in a prompt, you actually have to enforce it, this has got to be 100% compliant. And so, you know, the ways that we have approached AI guardrails is multifaceted. We have to stay one step ahead of malicious actors, we've got to stay one step ahead of where we think the legal and regulatory frameworks go for AI in general and then we got to be really attuned to the compliance environments that our customers operate in, you know, depending on what vertical they are, and you've got a lot of health care customers, a lot of financial services and insurance customers. And so, it has to be something that you care about every single day or else it just, you know, won't get implemented

as deeply as the enterprise requires.

[00:15:19] Theresa Tung Well, it's so new, right? So as an industry, we're figuring out the standard patterns in both, the areas of risk as well as the solutions, so indeed, I'm looking forward for 2024 as you are, as the year of security generative AI. May, continuing with you, right. You know, we often hear about Retrieval Augmented Generation, RAG, as a way for a company to apply their own first party data and I think Writer does much more than that. Can you share some of the components needed in that full stack GenAI application? And I think because of that, you know, there's a lot of both opportunities for, you know, securing and making it accurate, need to check its legal and making sure it's transparent.

[00:16:10] May Habib So, you know, Retrieval Augmented Generation is so new. If I had said a year ago on this podcast, you know, and when you think about RAG, people would be like, wait, what are we talking about? RAGs in AI? And so, it's awesome how much progress we've all made. I think what we are learning, you know, as well as I think the market is learning is, you know, taking all your data and sending it to, you know, a third party embeddings model only to get embeddings that you then have to send to a different third party vector DB that you then take into yet another third party LLM in the context of enterprise data that is messy. That is dynamic, that is ever changing. These bills are racking up fast and the accuracy results aren't there to pay for it. And so, what we said was, and we were doing this too early on, what we said is, all right, we need a different approach. These algorithms and this approach is just not getting us the accuracy rates that we need and a lot of it has to do with the way data is structured and the way vector databases work. And so, we've got this hybrid approach, it's a graph-based RAG and allows us to actually use Large Language Models. We have fine-tuned a model that all it does is create relationships between entities in enterprise data. And so, you get a think of it as like an AI semantic layer that helps the model much better understand and digest really dense enterprise data and rank results in that data that might better be suited and made better be candidates to answer a particular question. So, you know, RAG is, again, a multifaceted, improving the results. We've taken, you know, multiple approaches to improving results from, you know, that graph-based approach to something we call Retrieval Aware Compression, where we try to get a lot more understanding of the data into the model via some compression techniques that are pretty novel. And then, lastly, you know, we use a set of techniques called Fusion and Decoder, and this is, you know, something that we've done research on, communities' done research on, but you know, it's essentially a way to get the data that you use for RAG kind of closer to the transformer. And all those techniques are not easy and starting over for every use case is no small feat. Most folks don't have, you know, the engineering hands around the table for it and so part of our full stack approach is to make that tooling integrated into the way you build use cases in Writer and so it's a lot more repeatable for companies to kind of reuse over and over to build additional, you know, knowledge and retrieval based use cases.

[00:19:33] Theresa Tung And I think the lessons learned is Writer is clearly... It's literally your business to figure this out, right? But some of it could be for companies who are doing their own RAG models to look at your journey. Maybe they're starting with that semantic layer and using that domain knowledge graph alongside the Large Language Model as a means to help them increase the accuracy, increase the contextualization, even be able to catch the results and find relationships that they can govern and then reapply as opposed to, you know, really just working without that framework.

[00:20:12] May Habib Yeah, and we are planning to release an API to the graph-based RAG for folks to use inside of applications where we're going to start with, you know, folks who are using Writer already but then open it up to everybody.

[00:20:28] Theresa Tung That's great. So, Marc, coming back to you, you know, when introducing generative AI, one of the approaches that you talked about was how important it was for the human in maximizing value. So, what are some approaches that you took to make it easier for our humans at Accenture to embrace and benefit from the advancements?

[00:20:52] Marc Appel Yeah, it's really interesting to sort of talks to the change management portion of how we rolled it out and some of the observations that we've learned early and started to apply as we've gotten more and more traction. One of the techniques that I will say, aside from making sure that you've got a strong champion group that you can rely on...

[00:21:15] May Habib You need a Marc... He's been really humble.

[00:21:18] Marc Appel You need a...

[00:21:18] May Habib You need a Marc.

[00:21:19] Marc Appel You need a lot of Marcs. You know, and this goes for our Writer rollout as well as all other things that we're doing. You tend to find those champions in unexpected places. You do need senior buy in to get those, you know, the teams moving, but those champions, you know, they have to have a little bit of charisma, they have to demonstrate excitement, they have to, you have to feel it from them when they're talking to you and when it happens, you know, I've seen a little bit of like a magic moment happen in every meeting that I've ever shared, any tool, especially Writer, with and I'll give you an example, and the insight here is that you need to schedule play time with people, right? Every time I've done that 10 minutes, one feature, I come to them with a use case, and I go, I think this would work for you, and I'll show you the example here. I remember when Writer actually released a feature called Ask Writer, which was sort of their open source, not open source, excuse me, open form ChatGPT competitor, if you will, but better, and I called up one of our leaders in our reputation group, and I said, hey, I know it's 5:45 on a Monday and no one wants to stay late, but can you give me 10 minutes? I think there's somewhat of a use case that you want to do. I've got the solve for it. I think you're going to like it. Got on the phone with her, showed it to her in, like, 5 minutes, immediately, she brought in two more people to the call. The call went for 45 minutes, starting at 5:45 p.m. on a Monday. We literally spent the time like, pushing the boundaries of what the tool could do, right? Can it do this? Yes. Well, what about this? Well, not, not quite, but I bet we can cheat it if we did it this way, right? I mean, this conversation happens every single time. And it all comes from a ten-minute call with a champion who's, like, willing to go out there and say, I know your use case, and I think there's something I can do here for you. And what ends up happening is not only is it the best call of their day because they're super excited, again, more excited, but it helps them be more strategic because then what they do is they re-analyze their bids, their use case and go, well, should we be working this way? Maybe we can end back to that Good Morning, Accenture example, right? It's like, well, what if we didn't do it this way? What if we did it some other way where we didn't even we skipped these five steps that we normally go through and we just go to the in this case, the leader and have them just get on a phone call, we record it and then we upload the recap, upload it into the system, and it does the recap for us and we generates the email for us and we just provide a little editing at the end, instead of... I go, I do the thing, I get a review, I have an editor come in, I do it... It's like 16 steps in the middle, and all of a sudden, you're like, oh my God, we can reinvent how we're doing this whole thing. So, those little play time moments, that 10 minutes completely changed in this particular couple of use cases, how they think about, what they work on and how they work on it, which was fantastic. And I see this happening all the time. And so, what I encourage people to do is schedule the playtime moments. I call it playtime because it is actually fun, and it really gets to places. So yeah, that's a huge part of the change manager program is really, is really focusing on those playtime moments.

[00:24:50] Theresa Tung Well, thank you. Thank you both. I think in this trust chapter, you know, we've discovered what's needed to be enterprise ready, that it needs to be secure, accurate, legal, transparent, that it should be part of a responsible AI program and not just for GenAI, but thinking about how to handle data in AI enterprise scale. There's some AI guardrails and certifications that may be coming, but it's not really figured out yet. So, there's some steps that we should be able to start taking, though, right? Some of them might be, for example, adding that knowledge graph or semantic layer to make it, you know, more accurate and contextual with what we're doing. And then, Marc, I think, love the final point about that play time and because

we didn't really, you know, when it's play time, you're really reinventing that future, right? And you're giving that person, that user the control as to I'm going to learn how to use this in a way that augments my role. We're not talking about taking jobs away, right? It's actually automating the mundane. So, I think that that play time does a lot to assuage that and then also show the benefits of carrot and the stick. So, with that, I have one, just one last thought for each of you. What's your advice for companies looking at deciding, you know, is the value real and can we trust it?

[00:26:16] Marc Appel Because the people part of this is so important, right, for me and for why I pushed so hard months ago to get our team its speed really is everything, right? And building the GenAI muscle because it's reinventing how you work and what you work on, you need to know it. It's one of these things you actually have to get in there, not talk about it, you have to use it, right? And so, getting people to use it quickly and fast is going to separate your team, your function in your company from others. And it's going to only widen unless everybody starts talking about this and that we were at. But, you know, from the speed perspective, you know, it's like let's just get everybody as many people as we can enabled have them start moving. Right? Learn, build that muscle and let's pick off those use cases that we know we can do quickly and fail or succeed quickly. Because as soon as you do that, when you're doing this play times all of a sudden, the world opens up and you're like, I can do my everything differently, right? And so, if you don't start, that's the problem. And I think that inertia of just making sure everybody has it and that they can start is the big thing to get over income.

[00:27:34] Theresa Tung Great advice. May, any additional advice?

[00:27:40] May Habib Oh, my gosh. I love that so much. I want to triple underline it and add no notes. I think velocity of learning really is everything because it's only speeding up. And if you're sitting on the sidelines thinking it's going to fall into your lap or magically get worked into the products that you are using, your teams are using, that's really the mistake. What generative AI is enabling is teams like Marc's to reinvent the way they work and build their own truly, right? It's the layer, all the capabilities that are coming. This is the layer between everything. And the sooner that folks understand the speed is the disruption and that the magic is going to be in re architecting workflows AI first, I think the sooner people understand, oh, wait a minute, no, this isn't just going to get built into every tool I use because this is a completely new paradigm and only when you get messy in it you do really see what that means.

[00:28:46] Theresa Tung Well, thank you. So, mic drop moment from Marc and May. Get started. Don't wait. The value is real and for these patterns it is secure. We know how to do the basics, even though we're still evolving with the regulations. But we know how to make it trustworthy, secure, legal and transparent. Well, thank you so much. And please join us next time.

[00:29:11] Marc Appel Thanks for having us.

[00:29:12] May Habib Thank you, Theresa.