Deloitte.

AI360

From Generative AI Use Case Fatigue to Scaling AI Roadmaps

Rohan Gupta: Hello everyone, and welcome to another episode of AI360. This is the show where we bring you the latest in AI in 360 seconds or less. I'm here for episode 2, and I'm joined by my colleague Mohamad. Mohamad, do you want to introduce yourself?

Mohamad Said: Absolutely, thanks, Rohan. Good to be here today. So good morning! My name is Mohamad Said, and I'm a leader in Deloitte Consulting's AI and Insights practice. And I help organizations gain actual insights from their data with machine learning models. I develop and implement scalable enterprise AI solutions across organizations' value chain to drive new revenue, better manage risks, and improve productivity, and of the last year or so, we primarily focused on helping my clients develop and implement Generative AI capabilities.

Rohan Gupta: Fantastic. Well, you're the right person for this episode because we're here today to talk about scaling use cases in the enterprise and how that can sometimes be a challenge. So could you talk a little bit about why customers are sometimes stuck in POCs and not able to move to that production use case. What are some of the hurdles that you're seeing?

Mohamad Said: So as you can imagine, there is a lot of excitement in the marketplace, and rightfully so, to experiment with different use cases. And Generative AI is really going to transform the way many companies operate and the products and services they deliver. And in the last six months, we've seen many organizations go through use case ideation sessions, explore the art of the possible, and rethink about how they want to do business. However, while you might be successful at experimentation, there's quite a different ball game to scale and operationalize AI solutions. And also the technology itself is changing rapidly as we witness new releases of large language models in the market and also the emergence of open-source models.

So the biggest hurdle that I see is the ability to navigate this changing landscape and take advantage of the latest technology while also implementing strong guardrails to deliver safe and secure solutions that can enable multiple use cases. And the way to overcome this obstacle is to develop enterprise platform capabilities that enable multiple use cases while allowing you to swap between different models, but also applying the same guardrails that ensure models generate safe and trustworthy content.

Rohan Gupta: Makes sense. I love the idea of this platform. So as companies are thinking about use cases, how does one go about prioritizing them, and what are some of the common challenges you face along the way?

Rohan Gupta: Interesting. I didn't realize that there were so many options out there today. Could you break down the benefits and the drawbacks of building your own model or potentially buying one from a third-party?

Mohamad Said: So we're seeing many companies do really well with use case ideation and explore how to use Gen AI to enable their business in different ways. I've had one client, for example, with a portfolio of 55 use cases in just one function. However, with limited capacity and budgets, it's really difficult to implement each one of these use cases—and also the complexity to implement these use cases varies significantly.

So to overcome these challenges, what I typically do is I advise our clients to consider three key factors. One is what value does the use case bring to customers. This can include, for example, improved customer experience, a product feature, or a service that meets a specific need or new insights. Second is the impact that the use case can have on the business, and this can include new revenue or improved productivity or even cost reduction. And then third, how feasible it is to implement the use case. And this is where we look at complexity, data availability, and also the labor costs associated with the use case. So really triangulating these three factors can help organizations strategically focus their efforts on the high-value use cases that align with their strategic priorities.

Rohan Gupta: It makes a ton of sense. I love the three-point framework. It's very simple to understand. So you talked about the platform, and you talked about these use cases, so what are some of the core capabilities required to actually enable these in the enterprise?

Mohamad Said: Sure. So, you know, once the use cases have been prioritized, we typically go through an exercise of mapping the use cases to core capabilities that are required for implementation. For example, I had one client where we helped them prioritize 26 use cases. But when developing the capability matrix, we really narrowed down on just nine core capabilities required to implement.

Now, for example, if you're implementing a chat bot to help customers search for products on your website, you will need a conversational user interface. You'll also need a search capability and a fine-tuned large language model that can sit in a retrieval augmented-generation architecture with a vector database and also a moderation capability with a feedback loop.

Now those same core capabilities can be used for a cross-sell/up-sell use case or even another application, such as helping internal employees search internal documents and forums for knowledge sharing. So, as you can see, you'll have many synergies across different use cases. And as you add more use cases, the incremental cost for implementing another use case drops significantly over time because you'll have the capabilities already implemented for that use case.

Rohan Gupta: That makes sense, and I heard that you touched on retrieval augmented generation, which I know a lot of clients are asking us about so we'll have to come back for another episode to talk about that. So, Mo, thanks so much for that. I know you talked a lot about use cases, building a platform. Can you talk about bringing this all together? What does the implementation road map look like for executives, and what should they think about on day one?

Mohamad Said: The first step is really to select at least one large language model [LLM] provider and having a reference architecture that can be scaled across use cases, and they can manage different LLM instances. When creating the road map, I typically advise my clients to sequence the capabilities by building upon what they already have created during the experimentation phase and then prioritizing those that balance the number of use cases they can enable and the ROI these use cases can generate.

Another key consideration is the budget cycle of the business and tapping into any existing funds that kick-start initial build and setup. The road map needs to align with the vision of building a broader enterprise platform foundation and sequence in a way that generates frequent wins as use cases come live. And then most importantly, the timing of these wins need to align with the broader strategic plan and priorities of the business at the executive level.

Rohan Gupta: That makes sense. Mo, this has been fantastic. Thank you so much for your time, and we'll certainly have you back for another episode soon.

Mohamad Said: Absolutely a pleasure. Thanks, Rohan.

Visit the AI360 Podcast Episode Library
Deloitte.com/us/AI360