Deloitte.

AI360

Build or Buy

Foundation Models

Rohan Gupta: Hello, everyone, and welcome to AI360, a series that discusses emerging topics in enterprise AI in—you guessed it—360 seconds. I'm your host, Rohan, and today we're joined by Baris, who will share his thoughts on a question of growing importance: Should companies build their own AI models or buy one from a vendor? Baris, thanks for joining us.

Baris Sarer: Thanks, Rohan..

Rohan Gupta: Baris, could you start by just introducing yourself briefly and sharing your experience with Generative AI.

Baris Sarer: Sure. My name is Baris Sarer. I'm a principal with Deloitte Consulting. I lead our AI practice for telecom, media, entertainment, and technology industry as part of which I've been leading a growing number of high-impact, transformative Generative AI projects with our clients.

Rohan Gupta: Fantastic. Well, Baris, could you speak to what marketplace trends are you seeing that underline the criticality of this choice? Like, why should companies be building, or why should they be buying?

Baris Sarer: Great question. As our clients are starting to now think through how they build their Generative AI solutions stack, we're seeing four categories of tooling options emerging. One, large language models offered by leading cloud providers like Azure, AWS, and GCP.

Another one emerging is enterprise software with embedded Generative AI capabilities. We're seeing several large SaaS players already announcing their own GPTs or Generative AI solutions, especially for office productivity. And we're anticipating significant market action in terms of new products getting rolled out in that space third quarter of this year onward.

The third option is point solution, provided by third parties for use cases that require minimal customization. We're seeing a large number of third-party SaaS solutions emerging: synthetic voice, image generation, autonomous coding, and some more rudimentary forms of text generation would be the typical use cases.

And lastly, there's a sizable open-source, large language model and training data market shaping. For example, if you go to Hugging Face, which is a portal for models and data, one can find thousands of open-source models and data sets that can be used. While we think this will be an exception rather than a norm for a lot of our clients, we do expect to see some experimentation with training and fine-tuning open-source models.

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?

Baris Sarer: Well, you know, "build versus buy" is the perennial IT strategy question, and it applies to Generative AI as well. We view build-versus-buy decision as not binary, but rather a spectrum of options ranging from plug-and-play commercial off-the-shelf solutions at one far end of the spectrum to custom-built models at the opposite.

Now costs is easy to purchase and implement; however, for large enterprise purposes, most use cases that would drive significant business value would be too complex or too nuanced to make plug-and-play solutions a viable option. Similarly, if you go to the opposite end of the spectrum, building a large language model gives you full control on data and model behavior, which is a good thing. But on the other hand, while there are some enterprise and startup success stories out there with custom models building and, more importantly, maintaining a commercial-grade model is not a trivial exercise, and our clients need to keep that in mind. In the middle of the spectrum, our clients have the option of acquiring a private instance of an existing large language model supported by a large technology provider and fine-tuning it with their own data. This option offers better time to value and maintainability, but what you do it at the expense is control over the entire range of data that the model is trained on. We're seeing a great majority of use cases being delivered in this model.

Rohan Gupta: Very interesting. So, what are some of the factors then that businesses should consider as they go about making this critical decision?

Baris Sarer: Well, so first of all, this is not one decision. It's going to be a series of decisions, but to help our clients address that question, we developed a decision framework with four key considerations that you should apply to your use cases. One, you have to start with your business strategy. What is more important for you? Is it time to value? Is it monetization of your proprietary data? Is it differentiation, or is it creating a reputation for technology innovation to attract certain types of talent?

Second, you should consider your risk tolerance. For example, your tolerance for external bias in the data, your tolerance for liability exposure, hallucinations, model upkeep from a compliance standpoint—that would be a longer-term consideration. These should all factor into your decision.

And then thirdly, cost is always a critical factor. And it's not just about building or training the model but remember that whatever you buy, customize, or build, you're going to have to maintain either on your own or with your vendor partner. So, you have to apply that total cost of ownership lens to it.

And last, but not least, is your data. The key question in that dimension would be: can you create a differentiated solution or a path to a new revenue stream with your proprietary data? If the answer is yes, maybe you should start considering heavily customized large language models or something built from the scratch if you have the technical chops and financial means to do so. If the answer is no, you should probably go back to the commercial off-the-shelf side of the equation or something in between.

Rohan Gupta: That makes sense. Pretty interesting. I know we're still in the early days of Generative AI, so any advice for how anyone thinking about this can be future-proofing their decision considering how fast this market is moving?

Baris Sarer: I have to start with challenging that concept. First of all, given the pace of change in this space, it may not be even realistic to chase future-proofing. You may be tempted to pick winners and losers from a technology or vendor standpoint, but I'll submit that we're too early in the game to do that.

So as our clients build their Generative AI solutions stack, our recommendation would be to take a toolkit approach. Understanding, as I said before, build-versus-buy will not be a one-time decision, but you will make multiple such decisions, depending on a use case, your business objectives, which may evolve over time.

You should assume that over, you know, fast-forward two to four years, you're going to have a variety of solutions for a variety of use cases. And instead of trying to future-proof specific technology decisions, you should develop a North Star reference architecture with proper governance and guardrails, where all these different Generative AI solutions can work together and deliver business value in a

safe and secure environment.

Rohan Gupta: Makes sense. I guess some principles truly are evergreen. Well, Baris, thanks so much! You've been our first guest on AI360. Appreciate the time, and we look forward to having you back in the future.

Baris Sarer: My pleasure.

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