AI360 podcast

Season 1, Episode 5: Open source vs. closed source LLMs

Host:

Rohan Gupta, Principal, Deloitte Consulting

Guest:

Baris Sarer, Principal, TMT, Deloitte Consulting

Rohan Gupta: Hello everyone, and welcome to another episode of AI360. This is your favorite AI show that talks about the hottest topics in 360 seconds or less. I'm really excited to welcome back Baris in the show. Baris, do you want to quickly introduce yourself

Baris Sarer: Hi Rohan. So first of all, good to be back on AI360. My name is Baris Sarer. I'm a principal with Deloitte Consulting, and I lead our AI services for Telecom, Media, and Entertainment and Technology industries in the United States.

Rohan Gupta: Thanks, Baris. So I know that open-source and closed-source models has been a really hot topic, especially in the world of generative AI. So maybe let's start by talking about that. Could you help us define this? What is open source, and what is closed source?

Baris Sarer: Sure. Let me start with closed source first. In a nutshell, when enterprises purchase commercial software, they're purchasing a license to use it for an indefinite period of time. They can configure the software—to the extent that it's configurable—but they cannot change the source code, which continues to be owned by the software vendor. This is closed source. Open source is when you actually have access to the source code and you have the ability to modify it under some kind of open-source license, like new GPL, Apache, MIT, and others. Now, note that "open source" doesn't necessarily mean that the product is free of charge. A for-profit software company can release open-source software and then charge for support, additional features, etc. In other words, there are ways to monetize open-source software. And as a matter of fact, we have many examples of successfully monetizing open source through either revenue generation or M&A exits. Examples of which would be GitHub, Red Hat, Unix, MongoDB, etc. So it's a well-established marketplace for open source as well.

Rohan Gupta: Perfect. Thank you. So what's happening with generative AI—and with large language models [LLMs], in particular—as it relates to open source? Why are people talking about it so much?

Baris Sarer: Well, the open-source LLMs got really popular I would argue around early March this year with the leaking of an early iteration of Meta's LLaMA. Meta originally opened LLaMA source code to researchers only, then it got leaked and once it was leaked, Meta made the decision to make it broadly available. Then came the infamous Google memo early May, where some Google AI researchers claimed that it would be nearly impossible for closed-source models to create long-term competitive advantage over open source, which I tend to agree. And that the open-source model will outperform closed-source models, which remains to be seen. That certainly is not the case today. But at any rate, these developments attracted everybody's attention to this fledgling marketplace.

Now, we have many open-source models today. The leaders in this field—and there are, like I said, many of them—but LLaMA-2, by Meta, is truly popular. I heard that it got downloaded over 30 million times. And in a matter of months, Falcon by Tech Innovation Institute of the Arab Emirates, TWI, Mosaic MLs MPT models, Vicunia, and one of the early contributors to open source was actually OpenAI with Codex, their co-generation model. These models come in all sizes and shapes. You've got chat models, you've got foreign language specialized models, you've got models focused on coding, and yet they're also very small or smaller models like Mistral, for example, that are easier on your compute resources. It's a really fast-evolving space, with a significant proliferation of these different types of models. The Hugging Face website, which is a giant repository of open-source models and training data, is a good place to keep an eye on, and then Hugging Face ranks these models based on their performance and has become a popular reference site from a benchmarking standpoint as well.

Rohan Gupta: Yeah, that makes a ton of sense. In fact, a lot of clients I'm talking to as well frequently visit the Hugging Face benchmark page. So Baris, you know, you talked about the number of open-source models. It's just exploding. As a buyer and as a user, how do I think about open source versus closed source, and how do I make that decision?

Baris Sarer: Well, as always, you have to understand the trade-offs. Certain factors favor open source, such as a) could be cheaper, potentially. But you cannot simply assume that; you really have to do your math to ascertain it. And other potential advantages: It gives you some flexibility. You can get started right away, download it, tinker with it, tweak it. You don't have to go through a complex contractual arrangement. Thirdly, you're in a more dynamic and experimental market. You may gain access a lot faster to a technology or use case breakthrough. Another consideration is it gives you the ability—and I think is probably a more important aspect—it gives you the ability to heavily fine-tune the model. In other words, you have more control on the model, which can be critical for certain use cases. It also gives you some flexibility to deploy a solution across multiple cloud platforms or in a hybrid model or on-prem, although this is not necessarily a function of being open source. And lastly, since you own and operate the model, it cannot be changed on you like the managed closed-source models. But on the flip side, obviously there's no enterprise support. You can tap into a community of developers, or whatever information you can find online, but there won't be a formal support structure, at least not at this point in time. Also, there's the risk of open-source projects without proper financial means [that] would be more vulnerable to neglect and defects that could lead to, I don't know, security incidents, for example.

We had that in the past with [the] open SSL debacle a number of years ago. The Equifax data breach was also attributable to an open-source component. And then, given the constantly evolving nature of alarms, can an open-source code keep up in terms of performance and compliance? It's another important potential drawback, especially on that last point. They may not provide enough transparency with the training data, which could aggravate liability concerns. So these are some of the questions that you need to ask and make a well-informed choice.

Rohan Gupta: That makes sense. That actually seems very closely related to another conversation we had recently, which was around build versus buy.

Baris Sarer: Right.

Rohan Gupta: So as you're advising our clients, Baris, when should they consider an open-source model?

Baris Sarer: I would consider open-source solutions when one or more of the following conditions are met: One, you want to run that large language model on-prem, then you have more open source options, in that case. Two, you have the time, skills, and motivation to test and compare different options. Thirdly, you want to heavily fine-tune your own model, perhaps with a specific industry focus. You want to handpick the data sets and exert full control over training methods. Four, you want to create a specialty LLM and potentially monetize it. And with a really flexible, open-source licensing model, you could do that. And then, you've done your homework on the cost structure and found it to be more attractive than closed models—and when that really matters, open source is a good consideration. So these are the situations in which I would consider using open-source large language models.

Rohan Gupta: Well, Baris, that was incredibly insightful—another fantastic episode of the AI360. I might bother you with one additional question, just because I'm curious. I know open source has already been a hot topic in 2023, but as we think about, you know, looking ahead, what are some predictions you have for how open source is likely to evolve over the next six to 12 months?

Baris Sarer: Well, I predict that it will continue to grow in size and relevance. I also predict that we'll see more enterprise adoption. And to be sure, open-source models are already in use in many of our pilots with our clients. We'll see some of these go into production in the new year. And my final prediction is that within 12 months, we'll start seeing some successful examples of open-source monetization. The wild card is going to be AI regulations and how they may potentially impact the open-source space.

Rohan Gupta: That makes sense. Thank you so much. It sounds like we actually need to do a broader set of predictions that go beyond open source. Maybe we can put our heads together on that.

Baris Sarer: Yeah, I would love to.

Rohan Gupta: Well, thanks, everyone, for listening. This has been another episode. Please like and subscribe, and come back for our next episode that should be out very soon.

Visit the AI360 Podcast Episode Library to learn more.

This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to deloitte.com/about.

About Deloitte
Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.

9486477