

Speaker 1 ([00:00](#)):

Hey, everyone. We've made several big changes to how we provide podcasts to you, and of course we'd love to get your feedback. See the show notes for a link to a five question survey, it'll only take you a couple of minutes, and let us know what you think.

Jillian Forde ([00:14](#)):

This is the AWS Podcast. I'm your host for today, Jillian Forde. And if you want to learn generative AI this year, you want to learn the foundations, then stick around for this episode because there is an exciting course that is out now that you can take, and we're going to talk about this exciting course today from some of the makers of this course. So let's do some intros of these amazing two people that we have on the AWS Podcast, so Antje.

Antje Barth ([00:44](#)):

Yes, thanks so much for having me, Jillian. Antje Barth, I'm a principal developer advocate for generative AI here at AWS.

Speaker 1 ([00:52](#)):

Fantastic. And Tommy.

Tommy Nelson ([00:53](#)):

Hi. Yeah, also very grateful to be here. Thanks for the opportunity, Jillian. My name's Tommy Nelson, and I am a curriculum product manager at DeepLearning.AI.

Jillian Forde ([01:04](#)):

Fantastic. So we've got this course with AWS and DeepLearning.AI. So Tommy, For those who aren't familiar with DeepLearning.AI, what is it?

Tommy Nelson ([01:12](#)):

So DeepLearning.AI is an education technology company, founded by Andrew Ng, one of the pioneers of machine learning, not just in education but just period in the world, and we build courses that aim to grow and connect the global AI community. And so, we really believe that AI is the new electricity, this is something that's going to be permeating our lives, it's going to create so much opportunity for people, for communities. We want to give people the resources they need to learn these skills, incorporate them into their jobs, and make this a part of their future.

Jillian Forde ([01:50](#)):

I love it. I'm sure a lot of people are already seeing that now, especially with generative AI, and they're really excited for a way to get started. So tell people about this collaboration with AWS and DeepLearning.AI.

Antje Barth ([02:02](#)):

Yeah, so we're super excited to be working with DeepLearning.AI, and the amazing thing is that AWS and DeepLearning.AI, we share the same vision, really to educate the developers out there on AI through high quality course content, and we love preparing more individuals, as Tommy said, for the AI jobs of today and tomorrow. And we've actually been collaborating for a couple of years right now. So back in

2021, we released the Practical Data Science on AWS Cloud Specialization together, and this one was really to help developers learn how to build and scale data science projects in the cloud. And now, just a few weeks ago, we launched this amazing course to help engineers and data scientists to gain more foundational knowledge in the space of gen AI, and specifically around large language models.

Jillian Forde ([02:51](#)):

Well, this is a really good teaser then. So please tell us about this course. And I hear that there's already over 70,000 people that are enrolled.

Tommy Nelson ([03:00](#)):

It's been really exciting to see the response. The community of learners at DeepLearning.AI is large and global, we have hundreds of thousands, even more than a million learners all around the planet. And it's very clear from the way that they have responded to the launch of this course that there is a lot of hunger out there to learn how to use this technology, understand how it does these seemingly magical things that it does, and figure out what the opportunities are for them to incorporate into their work what opportunities there are to build these really great applications.

([03:32](#)):

The course is really designed to give very deep foundation of the fundamentals of this technology. We're focusing on practical applications, but we really want people to understand how the practical applications are made possible. So the course is broken down into three weeks of content, it's approximately 16 hours of video, featuring Antje and several other instructors from AWS. We also have some quizzes in there, and we have some practical hands-on labs that we've built in partnership with Vocareum that allow our learners to actually try out these technologies in a cloud setting, which is really important for generative AI. The models, the infrastructure needed to run them are large, it's not something that's super easy to get running on your own computer, so we've set up these labs that allow people to try this out in something approximating a production setting. So we really think that just taking this course over three weeks prepares you to understand how generative AI works, but also how it is going to incorporate into the decisions and strategies that your company, your business, is making right now.

Jillian Forde ([04:39](#)):

This is so exciting. So you said that the course is 16 hours, and I'm sure people here who are listening to this, they all have something in common, whether it's they want to change careers related to generative AI, maybe someone wants to grow in their career, maybe start a business that uses generative AI. So can you give them a sense of, after taking this course, again, that's 16 hours, what are the things that they're going to be able to do?

Antje Barth ([05:02](#)):

Yeah, absolutely. So as Tommy said, the goal of this course is really to equip you as a learner with the foundation and the foundational knowledge of how to work with generative AI and LLMs. So for example, a big focus of the course is around a concept that's called fine-tuning, and fine-tuning really helps you to customize the existing models with your specific data, and helps you to align those models for your specific use case. So we'll prepare you to really walk away from the course with a deeper intuition, how to use this new technology, and also with code in your hands so you can get started quickly with your own projects.

Tommy Nelson ([05:41](#)):

I think Antje made some really great points there, and I think this idea of building up that foundational knowledge is the really important piece here. This is a technology that is moving at lightning speed, and also, the products we're using out in the world that are making use of this technology are already very powerful. So I think we've all been deeply impressed by how some of these chatbots behave as you interact with them, and we really wanted to make sure that we were breaking that open, making sure that people understood what was happening under the hood, and also seeing how the various different pieces of this stack can fit together so that you actually are able to shape the way that your LLM behaves in production, and also how you can connect it up to other pieces of technology for your applications. You're really seeing the potential beyond these chat tasks, and understanding how this fits into these more realistic business-driven use cases.

Jillian Forde ([06:35](#)):

This is amazing, and I know, Antje, your explanation of it was so succinct. I think some people might think it's too simple, am I actually going to be able to do that much with it? But I can share from my own personal experience, so I work with early stage startups, and the exact thing that you saying that these learners will be able to do afterwards is what I'm seeing probably hundreds of startups are forming a business around. What you learn from this course could be just the premise to be able to get you one step forward of that business idea that maybe 10 years ago wasn't possible, that now is. So this is really exciting of what it is that both of you, and both of these teams, really have put together. So now, let's get into the details of this course. So what are some of the highlights going to be in each week?

Antje Barth ([07:23](#)):

Yes, so we cover a lot of concepts, super exciting. Let me just walk you briefly through the three weeks of content. So in the first week, you will see the transformer architecture that powers so many of those large language models, and we will discuss how the models are actually trained, and also talk about the compute resources you need to develop them. But you also explore how to guide the model output at inference time, this is called prompt engineering. So we are spending a fairly good amount of time to teach you the basics there, and also discussing the specific configuration settings you can use to guide those models into the outputs you want to have.

([08:04](#)):

And then, in the second week, we really dive a little bit deeper into how to adapt those pre-trained models to your specific tasks and data sets, what I said earlier, through a process called fine-tuning. So we're going to discuss instruction fine-tuning, single task, multitask, fine-tuning, super exciting. There's also a new variant end of fine-tuning called parameter efficient fine-tuning, or short path, that helps you fine tune those very, very large models that have billions of parameters using much smaller resources, sometimes even a single GPU, so you'll learn all about that. And then, we also dive into the metrics and actually help you answer the question, "How do I actually evaluate those models and the performance of LLMs?" Which is not that easy in this space.

([08:51](#)):

And then, in the third week, we actually discuss how to make the responses of the models more human-like, and how you can align them with what we call the human preferences. There's a technique very popular right now called reinforcement learning from human feedback, short, RHF, and we are going to discuss RHF really in depth. And it's usually around how to increase the helpfulness of models, make sure it's honest in the responses, and also the responses are harmless, so the three Hs. And you'll also have a chance in each week to get hands-on with these techniques in our lab environments.

(09:31):

And in the last week, we're also touching on some of the more advanced topics, like retrieval augmented generation and advanced prompting techniques. You might have heard about chain of thought prompting, or react. So we're going to explain all those concepts, show you a lot of examples, and what they do, basically, those techniques is help to integrate your large language model with custom data sources, or APIs, to even improve the model response further, and give it ability to interact with the broader world.

Tommy Nelson (10:00):

I think one of the things that was really fun about developing this course was that we got to have some really great conversations with industry leading experts, people at the research teams, AWS and other places, and we really have been bringing things hot off the presses from these research teams into the course. So we did a quick check, and the median publication date of the papers we discuss is late 2022. This is very recent research that's already making it into production, and we want to help the learners understand what these advanced techniques are, and the impacts that they're having on making these models more useful in production and deployment. So that was a really fun experience, taking that deep dive, really looking at these concepts as they're emerging from the research teams.

Jillian Forde (10:48):

I love that you just said that, because I know when a lot of people are taking anything tech related, they're always wondering, "Am I going to be able to use this in the next five years? Is this something that maybe is no longer going to be relevant?" Or, "Am I learning old material?" So that's really cool that it's something that's really relevant that you're calling out in these papers. And now, let's get into, there's people here who are like, "All right, what are some of the prerequisites that maybe I need to have in order to really get the most out of the course?" So let's go through that.

Antje Barth (11:17):

Ideally, you're an engineer or a data scientist, and you want to learn the technical foundations of generative AI, you want to learn about the concepts that power those models. And it's helpful if you have some familiarity with Python or the foundations of machine learning.

Jillian Forde (11:34):

So Tommy, what about people who are new to machine learning? Is this a course that could be a great course for them to get started with?

Tommy Nelson (11:42):

Yeah, absolutely. I think it's really important to know that, while, obviously, some prerequisites are useful, I think there's a lot in this course that people are just going to be super interested to learn. We've definitely designed the content, how we tell the stories, in a way that's accessible to a large range of people. I personally am not a machine learning engineer, have never been, my background is in astrophysics, so while I had lower level of familiarity with the specifics of this technology, I had a little bit of a background in the basics of machine learning, but I just learned an absolute ton as we worked on this course. We used to joke in our team meetings that I was learner zero. If I wasn't following along, if I couldn't understand, it meant we had to work harder to get the content right to tell the story in a more accessible way.

(12:29):

But yeah, I think anyone who is interested in the big picture of what this technology is, but also wants to know the details, wants to be able to talk to someone in their engineering team, or to follow some of the podcasts like this one, the other things that AWS produces, or even read along with The Batch, which is DeepLearning.AI's newsletter that we put out every week, which is very much for a developer audience, I think this course gets you up and running very quickly with what you need to do. And I've loved following the conversations happening out in the world, both in business, in the technical community, following them in more depth. I really feel like I understand what's going on there.

Jillian Forde (13:08):

That's a great call out, because I know there's a lot of people who are probably listening here, and they know that they see generative AI as so important, but they know that they're not the person that's going to go out and actually build it themselves, but they want to be able to have conversations about it, they want to be able to understand what's going on for their career. So that's great that you're calling out that this too is a course that they can use to at least have a better understanding.

Tommy Nelson (13:33):

The other thing that I would say is we have the practical experience here, and then, at DeepLearning.AI, we're also building out short courses that are designed to give people an hour, an hour and a half, experience, where they can try some of these generative AI tools for themselves. And I think the pairing of our long form course that we've built with Antje and her colleagues is a really nice setup to then go off and also try these practical hands-on experiences that are available on our platform, from AWS and from other educators out in the world.

Jillian Forde (14:05):

So Tommy, I have a question for you, because I know a lot of people, they're thinking there's a lot of different courses out there related to generative AI, so I'd love to get your opinion of what do you think makes this course unique?

Tommy Nelson (14:17):

Yeah, that's a great question, Jillian, thank you. I think one of the things we want to do at DeepLearning.AI is create comprehensive thorough learning experiences for people, so that you are getting a complete picture of what's going on and are building that foundational knowledge and intuition that you need to not just go out in the world and download a notebook and try it out, but to be able to follow along, understand what's going on, and be able to debug when things aren't working. I think there's a lot of great educational material out in the world right now about large language models and generative AI, but it's still a very new field and there hasn't been that much time to pull it all together into a comprehensive course, so I think that's the approach we took. We wanted to put something together that walked people through the entire process of this stack for model training and optimization, all the way to incorporating this into real world applications and deployment.

(15:11):

Now, obviously, we're not going into super detail about this specific library or this implementation of training, we want to keep the conversation at a higher level than that so people understand the core concepts. But I think by the time you've taken the course, you'll have a really good understanding of how everything is working, and it's going to help you make better decisions about what solutions you need. I think that this is a really nice compliment to other educational efforts that are happening at

AWS, at DeepLearning, we're building this new short course platform where people can try a lot of these technologies out for themselves, and this course that Antje and her colleagues worked with us on is a wonderful start to being able to hit the ground running with those. I really hope that this is going to stand for a while as a go-to foundations course that will then let you absorb all these other great materials and offerings, the events that AWS does, and then get building and know why you're doing what you do.

Jillian Forde ([16:10](#)):

I love it. Oh, I'm so excited for all the listeners here who are going to get started, but I've got one quick question for each of you before we wrap up. So is there anything else that you would like to share that could be advice for any of the learners before they actually start the course, or anything else that comes to mind?

Antje Barth ([16:28](#)):

Yeah, so from my side, I would just say stay curious, even if you're not super familiar with machine learning, with generative AI yet, be curious, have a look peek into the course, and also, if you're in the space, I think there's great learnings. And a big shout-out to all of the over 70,000 people who already did that and jumped into the course. And hopefully, we could get some of you listeners here a little bit more excited, maybe if you haven't done it yet, and give the course a try.

([16:55](#)):

I also want to finish with a big shout-out to my colleagues and co-instructors on the course, Chris Rackley, Shelbee Eigenbrode, and Mike Chambers, who really helped to develop and deliver this course. And also, we developed the course, what we just briefly mentioned before, with various industry experts and applied scientists here at Amazon, at AWS, also at Hugging Face, so just a quick thank you here, call out to E.K., and nationally, Heiko, Sally and Philip. Thank you so much, this is just a few of the names. And then, of course, a very big thank you to the whole team at DeepLearning.AI, especially to Dr. Andrew Ng, who provided us with great guidance, and, of course, Tommy here, who made sure we're delivering at the highest possible standard, and also made sure we're not missing out on the fun creating this course.

Tommy Nelson ([17:39](#)):

Thank you, Antje. I think that's a really important point to finish on. I think compared to some other machine learning and AI tasks and learning experiences over the last few years, there's something intrinsically fun about language models. You can actually have a good time experimenting as you're working with them. We had a lot of fun sometimes asking models to help us understand concepts in the voice of some of our favorite characters, and Antje and I share a love of the television show Gossip Girl, who, as I'm sure learners who are familiar with the show know, has a very distinctive way of speaking, and it was just sometimes a really nice way of breaking up the long days to ask Gossip Girl to explain RLHF. And when you hear Gossip Girl tell Upper East Siders that the new big deal in town is RLHF, and everyone has just got to stop what they're doing and learn it, you know that you're onto something, that you're going to have a good time.

Antje Barth ([18:35](#)):

It was amazing.

Jillian Forde ([18:37](#)):

Well, I hope someone builds an app that uses Gossip Girl for, I don't know, maybe you generate songs related to this, or something, I don't know. But I can't wait to hear what other people build. So the course, tell people again where it's at, where they can go sign up, and that's really about it.

Tommy Nelson ([18:56](#)):

Yeah, so the core is listed on the DeepLearning.AI website, deeplearning.ai. And the course is delivered through Coursera, so you can also find the course through the Coursera search bar, if you search for generative AI with large language models, or for the names of any of our instructors, including Antje Barth, those will show up.

Antje Barth ([19:14](#)):

And we have a short link, so if you struggle finding it, you can also type [inaudible 00:19:19].li, and then /gllm.

Jillian Forde ([19:22](#)):

Fantastic. Well, I'm so excited for everyone here to go and get started with the course, I'm definitely going to take it myself. Thank you both for being here on the AWS Podcast.

Antje Barth ([19:31](#)):

Thanks for having us.

Tommy Nelson ([19:32](#)):

Yes, thank you. It's been awesome.